An Investigation and Review of Futility Analysis Methods in Phase III Oncology Trials.

by

Chad Winch

A thesis submitted to the Graduate Program in Epidemiology

in conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

(December, 2012)

ABSTRACT:

The general objective of this thesis was to improve understandings of design, conduct and analysis of randomized controlled trials (RCTs). The specific objective was to evaluate the methodological and statistical principles associated with conducting analyses of futility, a component of interim analysis, as part of the conduct of RCTs. This objective was addressed by first performing a systematic review, which included a detailed literature search, as well as data from a cohort of previously extracted studies. The systematic review was designed to identify futility analysis principles and methodologies in order to inform the design and conduct of retrospective futility analyses of two completed NCIC CTG trials. The results of these trials have been previously published; one trial met its stated endpoint and the other did not. Neither trial underwent an interim analysis of futility during its conduct. The retrospective futility analyses assessed the accuracy of frequently used methods, by comparing the results of each method to each other and to the original final analysis results. These assessments were performed at selected time points using both aggressive and conservative stopping rules. In order to increase the robustness of the comparisons, bootstrapping methods were applied. The results of this thesis demonstrate principles associated with the conduct of futility analyses and provide a basis for hypotheses-testing of optimum methodologies and their associated trade-offs.

TABLE OF CONTENTS

LIST of TABLES:

LIST of FIGURES:

LIST of ABBREVIATIONS:

NCIC CTG	NCIC Clinical Trials Group
RCT	Randomized Clinical Trial
DSMC	Data Safety Monitoring Committee
IA	Interim Analysis
FA	Futility Analysis
HR	Hazard Ratio
CI	Confidence Interval
P	p-value
CR	Complete Response
PR	Partial Response
CRC	Colorectal cancer
NSCLC	Non-Small Cell lung Cancer
GI	Gastro-Intestinal
GU	Gynecological
DFS	Disease Free Survival
PFS	Progression Free Survival
OS	Overall Survival
HSREB	Queens Research Ethics Board
CP	Conditional Probability
H_a	Alternative Hypothesis
H_o	Null Hypothesis
SAS	Statistical Analysis System
CALGB	Cancer and Leukemia Group B

EXECUTIVE SUMMARY:

A randomized clinical trial (RCT) is an important tool for evaluating differences in treatment effect. These trials are able to provide strong evidence to support a hypothesis or research finding by effectively limiting bias from both known and unknown confounders by randomly allocating (randomizing) patients to each arm of the trial. In order to ensure continued patient safety and study efficiency, an RCT will often incorporate ongoing monitoring and assessment of data at pre-defined time points into their statistical analysis plans. These ongoing monitoring assessments, called interim analyses, are reviewed confidentially by an independent body, called a Data Safety Monitoring Committee, who then determines whether a trial should be permitted to continue or stop early. An interim analysis often includes an analysis of futility as one of its components. A futility analysis is designed to investigate the probability that the trial will not be able to achieve or demonstrate its objectives (i.e. how likely it will be, given the data available, that the null hypothesis will not be rejected). The ability to accurately and effectively conduct interim analyses of futility is important in randomized phase III clinical trials, because conclusive futility analysis results allow for scientific research questions to be answered earlier in the course of the trial. With early stopping of an RCT due to futility, results are more promptly disseminated, patient risk is minimized, and resources that would have been needed for further trial conduct are spared. The purpose of this study was to evaluate the methodological and statistical principles associated with conducting futility analyses by incorporating the results of a systematic review, designed to identify futility analysis principles and methodologies, to evaluate the conduct of retrospective futility analyses in two completed NCIC Clinical Trials Group (NCIC CTG) trials. The systematic review to identify trials stopped for futility was derived from a search of Ovid MEDLINE and EMBASE databases and articles extracted from two database cohorts of previous analyses assessing RCTs. The review determined that, although futility analysis methods are underreported in the literature, the two most commonly used methods to assess futility were *testing the alternative hypothesis at a very low significance level* and *stochastic curtailment based on conditional probability*. The trials selected for retrospective futility analysis using these two methods were Colon.17 (CO.17) because its final analysis demonstrated significant results (but

the actual absolute benefit of these significant findings was marginal), and Ovary.16 (OV.16) because its final analysis did not detect a significant difference between treatment arms, and resulted in a hazard ratio very close to 1. The retrospective futility analysis was applied at three time points in each trial, using both aggressive and conservative stopping rules for each method. The results were then compared using Chi square and Fisher exact tests. Bootstrapping methods were also used for each retrospective futility analysis, in order to limit improbable results based on chance alone. The results of the retrospective analyses indicated that while both methods became more accurate at reporting the outcome of the trial as more information became available, the more accurate and consistent method was stochastic curtailment. This method was able to detect that the trial would not be able to achieve its objectives over 90% of the time with 50% and 75% of the required events available. These retrospective analyses also showed that results of both methods differed substantially depending on the stopping guidelines used (i.e. conservative or aggressive settings). These findings are important to guide future research for investigating futility analysis methods, and provide important insight into how best to design and plan futility analyses as part of statistical analysis plans of randomized clinical trials in a phase III oncology setting.

CHAPTER 1 – Introduction:

1.1 Introduction and Background Information:

1.1.1 Randomized Clinical Trials:

A randomized clinical trial (RCT) is a very important tool for evaluating differences in treatment effect. A properly structured RCT is generally considered to provide the best evidence when testing a hypothesis, or in support of a research finding (Webb 2005). This is because with random allocation (randomization), the risk of bias is reduced by controlling for both known and unknown confounders. The aim of an RCT is to test a hypothesis in an environment where the patient characteristics in each group do not differ substantially. If this is done successfully, any differences in outcomes between groups can be reasonably attributed to the treatment/exposure in question. RCTs are designed by selecting a target population, and determining a significant sample from this population for testing. Eligibility of potential patients is determined by pre-defined characteristics, and subjects are randomized to an arm of the study. Specific data are then collected, and trial outcomes are subsequently evaluated. At the pre-defined 'end' of a trial, a final analysis of the accumulated data is undertaken, and the objectives of the trial are assessed. The conclusions are then disseminated to the scientific community. For a number of RCTs, the data are sometimes analyzed at specific time points throughout the course of the clinical trial. These mid-trial analyses, called interim analyses, provide a type of 'snapshot' of clinical data that can be used to provide insight into the feasibility of a trial, and to identify and act on extreme results (both positive and negative results). The information submitted for an interim analysis is provided in a confidential manner, and is used for the purposes of ongoing oversight and monitoring of an RCT by an independent panel of specialists, called a Data Safety Monitoring Committee (DSMC) (Bassler 2008). A DSMC uses the data provided for interim analysis to assess the integrity of an ongoing RCT, and to make decisions as to whether or not it remains ethical to continue a trial.

1.1.2 Oversight of Clinical Trials – Data Safety Monitoring Committees (DSMC):

The structuring and implementation of Data Safety Monitoring Committees is a recent occurrence in randomized clinical trials. DSMCs began to oversee clinical trials by means of interim analyses in the 1970's, and this process soon became standard in the late 1980's (Jennison 2000). Numerous papers were published

outlining the principles of these monitoring plans, and their implementation into trial design (Geller 1987, Green 1987). Since then, DSMCs have been empowered with the ability to decide whether a trial should be allowed to continue, or should be terminated early, based on the confidential results of interim analyses. Not surprisingly, a majority of funding, ethics, and regulatory bodies now consider independent monitoring with DSMCs an essential and integral part of any major randomized clinical trial (Bassler 2008). For example, the Food and Drug Administration provides guidance for sponsors of clinical trials on the establishment and operation of DSMCs (Anonymous 2001). Because of this, it is not uncommon for oncology groups or research organizations (any potential sponsor) to stipulate that they would not join an international trial if an independent DSMC is not available for that trial. Although the DSMC holds the power to decide whether a trial should continue, and are provided with interim analysis data to inform those choices, there are conflicting opinions that exist in the literature regarding decisions to stop trials early based on interim analysis results. In order to comprehend the importance of interim analyses and their role in the monitoring and oversight of clinical trials, it is important to describe the findings provided at these analyses, and the potential consequences of this data.

1.1.3 Interim Analysis:

The ultimate goal of any RCT is to provide evidence for the underlying research question (i.e. reject the null hypothesis, and accept the alternate hypothesis) in a manner that is safe for patients and efficient for the sponsor and other organizations involved in the study. This research question is designed to provide evidence of patient benefit, whether that is to show efficacy of an experimental therapy, or a more favourable toxicity profile of a treatment. To accomplish this endeavor, a sound methodology and statistical analysis plan is developed and implemented, and a DSMC is used to monitor a trial's progress in order to assess if truncation, or premature study closure, is warranted. This ongoing monitoring and assessment of data, according to the analysis plan, at pre-defined points along the course of a clinical trial is the basis of an interim analysis (IA). The DSMC is provided data in order to make informed decisions on the ethical and safety implications associated with allowing the trial to continue, versus recommending an early trial termination. There are potential consequences associated with either decision.

In general, reasons for early study closure on the basis of interim analysis include poor accrual, unexpected and often severe toxicity, external information, futility or lack of efficacy, and a significantly large difference in efficacy between study arms (Grant 2004, Mueller 2007). An interim analysis, therefore, will include any number of these components, but may not include all of them. A DSMC will investigate these reasons based on the data provided, and if their findings are troubling enough, they may recommend early closure of the trial.

There are advantages associated with stopping clinical trials early. Early closure can prevent additional patients from being exposed to ineffective or toxic therapies (i.e. uphold the safety and best interests of current and future patients), whereas positive results of truncated trials, which may potentially alter clinical practice, can be made available to the public in a timely manner. This means by successfully closing a trial early, the scientific question is answered sooner, the results are shared with the scientific community faster, and the safety of current and future patients is upheld most efficiently. In addition to the benefits associated with the safety of the patients, and with information dissemination, closing a trial early can also be beneficial for cost recovery and resource management of the individuals and organizations running the trials. Clinical trials require considerable resources to operate, and these resources may be re-allocated once a decision is made to terminate. It also potentially allows a new drug to be introduced to the market earlier if superiority is demonstrated.

Unfortunately, however, there are also some potential disadvantages that exist when deciding to close an RCT early on the basis of interim findings. The major limitation to interim analysis data is that, because the information is provided early in the course of a trial, there is a greater potential for error. Data provided at interim analysis are premature, and premature data can be unstable and potentially susceptible to extreme and anomalous results. This variability and susceptibility to the effects of chance alone can result in a presentation of inaccurate findings. To further complicate this problem, early release of interim analysis findings can also corrupt the conduct of the trial itself. This is especially true when the trial is still accruing. The new information about a more beneficial arm (whether that benefit may be efficacy or minimal toxicity, etc.) could result in increased crossover between treatment arms, a change in follow up, or a change in the characteristics of the patients enrolled. These changes can influence the final outcome of the trial and its long term results, and can further promote inaccurate findings. Dissemination of inaccurate findings can have similar consequences on

patient safety as allowing a trial with a less effective and potentially more toxic experimental therapy to continue, and there lies the issue with interim analysis and decisions to close a trial early. This concern for the variability of interim data, and the risks and benefits associated with early trial closure, has been a focus of investigation for trials that have stopped early for reasons of benefit (Mueller 2007).

1.1.4 Interim Findings of Superiority – Early Closure for Benefit:

There has been an increasing trend of early termination for benefit in randomized clinical trials. A recent systematic review found that these occurrences have more than doubled since 1990 (Montori 2005). This trend has prompted a number of articles to be published expressing concern and skepticism of the results of trials stopped early for benefit based on interim analysis results. These articles, some of them large scale systemic reviews, claim that such trials often tend to overestimate treatment effects, have less than optimal power due to a lower number of events, and often fail to adequately report relevant information about the reasons for stopping the trial (Trotta 2004, Goodman 2007, Bassler 2010). As a result, the literature expresses a concern for the implications of truncated trials in clinical practice, and warns clinicians to be skeptical about these potentially misleading results.

There are many examples in the literature to this effect. The United Kingdom Medical Research Councils AML12 in acute myeloid leukemia found a significant survival advantage for the experimental arm at interim analysis (HR = 0.55, 95% CI = 0.38-0.80; P = 0.002). The conclusion after deliberation from the DSMC was to continue the trial, as stopping boundaries were not specifically outlined, and to re-assess at a later time. Once a larger sample size had been accrued, this significant survival advantage could no longer be determined (HR = 1.09, 95% CI = 0.87-1.37; P = 0.4) (Wheatley 2003). Another well-known example is the Cancer and Leukemia Group B (CALGB) 9633 trial, whose significant survival advantage for adjuvant chemotherapy in resected stage IB non-small cell lung cancer resulted in early closure and the release of interim analysis findings (Strauss 2004). These findings led to the adoption of this practice into clinical guidelines. However, when later follow-up data were available, the benefit was no longer significant (McNeil 2006). These are all examples of the potential instability and volatility of interim data.

Despite these concerns, however, there have been a number of RCTs that have successfully been truncated for reasons of benefit, or superiority of the experimental arm, and have so far withstood the test of time (Goss 2005). In these cases, the scientific community has been alerted to these findings sooner, and patients have benefited from faster access to these therapies.

Another component of interim analysis, however, which provides a different perspective on the effect of the experimental arm is called a futility analysis. There has been substantially less investigation into this type of analysis in the literature, but it is commonly included in statistical analysis plans for interim monitoring of clinical trials (Friedlin 2009).

1.1.5 Futility Analysis:

In contrast to an analysis of benefit, a futility analysis (FA) is the flip side of the 'interim analysis' coin. Where an analysis of benefit looks to determine if the experimental arm is superior to the control (i.e. how significantly does the data favour the alternate hypothesis), an analysis of futility aims to determine the degree of equivalency or inferiority of the experimental arm to the control (i.e. how significantly do the data favour the null hypothesis). In other words, 'futility' is used to describe the extent to which a clinical trial is unable to achieve or demonstrate its objectives (Snapinn 2006), and looks for a lack of efficacy in the interim trial findings. It is important to note that these analyses of futility are performed in 'superiority trials', as opposed to trials looking to establish equivalency. A superiority trial is designed to investigate whether the experimental arm(s) of a trial is significantly better than the current standard or control arm, whereas an equivalency trial is designed to show that two (or more) therapies are not significantly different in their efficacy. For the purposes of this paper, the focus will be only on RCTs investigating superiority.

As mentioned previously, a DSMC may decide to close a trial early based on interim findings for reasons other than efficacy. These include poor accrual, unexpected and often severe toxicity, and implications of external information (i.e. results of other clinical trials or analyses, new therapies on the market, etc.) (Bassler 2008). Since the focus of this paper will be early termination for lack of efficacy, trials closed early for reasons such as study closure for poor accrual, toxicity, and external information from other sources will not be

described further. It is important to note, however, that an interim analysis may include any number or combination of these components, and the statistical methodology used for these components may differ.

1.1.6 Statistical Principles for Interim Analysis – an Overview:

An interim analysis of efficacy is a very complex entity, and trial decisions/recommendations based on these results must account for a number of factors including statistical, ethical, environmental, and external considerations. These analyses vary considerably, and several different grouped sequential methods have been proposed and evaluated (Skoylund 1999, Lachin 2005, Snapinn 2006, Lachin 2009), and new variations to existing methods are continually being investigated (Goldman 2008, Zhang 2010). The majority of these methods, however, involve similar guiding principles. First of all, they often involve a complex calculation of stringent statistical significance levels for each interim investigation, so that the overall significance level of a trial remains the same. This is because without adjusting the nominal significance levels at each interim analysis, probability of committing a type I error (detecting a false positive, or rejecting the null hypothesis incorrectly) increases when investigating for benefit or efficacy of the treatment arm. A similar adjustment is made when investigating lack of efficacy as well in order to reduce the chance of committing a type II error (detecting a false negative, or rejecting the alternate hypothesis incorrectly) at the interim. The result is that a boundary, or threshold, is created at each interim analysis which guides a decision to close a trial early. This boundary may be one or two sided, depending on the nature of the investigation, or may simply be a pre-stated rule (e.g. 'stop the trial if a complete or partial response –CR or PR– is not seen after 24 weeks in the first xx patients enrolled' (Sparano 1993, Blay 2007)). The interim analysis may also involve a calculation of a value designed to indicate the probability that the final analysis of a study will be statistically significant given the data available at this point. An example is the calculation of a conditional power, although other calculations of significance levels involving some threshold, rule, or barrier also exist (Snapinn 2006, Friedlin 2009).

Another important principle with interim analyses is that the boundaries used to indicate whether a trial should stop or continue can be set using aggressive or conservative methods. In other words, boundaries can include rules or guidelines that would result in a greater potential to stop a trial early (i.e. aggressive stopping

rules), or they can be set at a level that would make it harder to stop the trial early (i.e. conservative stopping rules) (Friedlin 2002).

The methods and timing of interim analysis may vary between trials, but the interim analysis plans are usually determined before a trial is activated, and are outlined and described in each protocol. These individual significance level and conditional power calculations are performed on the primary endpoint of the trial, and are the information provided to DSMCs to assess the efficacy of a trial. The aggressive or conservative stopping boundaries are pre-defined, but can vary from trial to trial, or even from interim analysis to interim analysis. Ultimately, it is the responsibility of the DSMC to assess these interim trial results, weigh the implications of advocating a decision to stop, and provide recommendations accordingly (See Appendix III for an example of stopping boundary guidelines based on the statistics described in this section). In each instance the question remains the same: is there scientific evidence to suggest that the trials underlying theoretical question merits further investigation, and does the decision to continue the trial remain ethical, safe, and in the best interest of the subjects included in the trial and future patients?

1.2 Rationale:

The focus of the first section of this thesis will be to conduct a systematic review of the literature in order to investigate the methodologies and reporting quality of futility analysis in clinical practice, and to describe the methods used to perform this analysis. The purpose of this review is to explore the nature of clinical trials that have closed early based on interim findings of futility, to gain insight into the methodology and setting of the phase III clinical trials that have closed for reasons of futility, and to use the results of these findings to inform the design and conduct of retrospective futility analyses described later in this paper. These retrospective futility analyses will involve applying the findings of the review in two NCIC Clinical Trials Group (NCIC CTG) led clinical trials that have previously published primary final analysis results. These retrospective analyses will be applied using both aggressive and conservative stopping rules, and will be compared with each other, as well as with the final analysis results of each trial, in each setting. The general objective of this thesis is to improve understandings of design, conduct and analysis of randomized controlled trials (RCTs). The specific

objective of the study is to evaluate the methodological and statistical principles associated with conducting analyses of futility, as a component of interim analysis, as part of the conduct of RCTs. In doing so, it is hoped that this study will provide insight into the effectiveness and accuracy of futility analysis, will promote consistency in clinical practice with interim analysis planning, will be able to provide a basis for hypotheses-testing of optimum methodologies and their associated trade-offs.

Although, as described above, the merits and disadvantages of terminating trials early for benefit have been extensively researched, there is little information available in the literature regarding the use of futility analysis. Certainly futility analysis is commonly used in the conduct of clinical trials (Friedlin 2009). Most articles discussing stopping clinical trials early, however, focus on stopping for reasons of benefit. As result of this focus, the information available on the current practice of planned monitoring of clinical trials with respect to futility analysis remains scarce. It is apparent in the literature that published reports of studies rarely include details on the strategies and details of interim analysis, and this is even more apparent when there were no 'significant findings' at the interim for trials that reach final analysis. A recently published study demonstrates this point efficiently. A survey conducted by the Italian National Monitoring Centre for Clinical Trials on the protocols of cancer clinical trials in the Italian registry from 2000 to 2005 determined that only 70.7% of protocols incorporated a statistical analysis plan for interim analysis, and only 56% also had a DSMC and were considered adequately planned (Floriani 2008). The surveyors concluded that insufficient attention is paid to the implementation and documentation of interim analysis. This same trend was echoed during data abstraction methods for this current investigation, which did not have access to the protocols themselves (the data extraction was limited to published articles and citations). This limitation is described in more detail in the Potential Confounders section of Chapter 2 (section 2.3.6).

It is also evident that the conduct and boundaries of futility analyses vary depending on the organizations involved in the clinical trial, the type of trial design, the investigation itself, and even the research endpoints (Friedlin 2009). Depending on the nature of the trial and the sponsor conducting the trial, the stopping boundaries can be set to be aggressive or conservative (if boundaries are set at all), and the scheduling of interim analysis varies as well. More information is clearly needed on proper guidance for establishing futility analysis,

as is more information on trends in the literature for this type of analysis, in hopes of establishing consistency in practice.

Finally, more information is required on proper futility analysis techniques because of the potential implications on the stakeholders associated with stopping a clinical trial early due to interim findings. If a trial investigating an agent or therapy that is most likely to be found inefficacious is allowed to continue, unnecessary risk may be placed on the patients. Cancer therapies often are associated with adverse effects and toxicity. If continuing a study of a novel therapy is allowed because of inaccurate or overly conservative futility methods, then a patient may be receiving a treatment that is potentially more toxic for very little scientific benefit. A worst case scenario would be when patients are allowed to receive a novel therapy that is actually inferior to the control or standard of care (i.e. the treatment arm is less effective than the control/standard of care). In addition to this, allowing a 'futile' trial to continue also has implications for the sponsor and drug company as well. Clinical trials cost a significant amount of money to run, and occupy significant resources as well. This is especially true with large scale phase III clinical trials in oncology, which tend to have large sample sizes, multiple sites, and a number of personnel required to oversee the project. By allowing a trial to continue when it is not likely it will be able to reach its objectives (or essentially has the answer already), the sponsor or industry company is subjected to large and arguably unnecessary cost associated to running the trial. In these difficult financial times, this can be a very big concern for cooperative groups, industry, and even government funding agencies. By accurately stopping a trial early, one can limit the risk imposed on patients, minimize financial investment, and reallocate resources to new innovative projects.

An alternate argument, however, is that there are also some potential disadvantages to closing a trial early or without enough evidence. While it seems appealing to stop a trial early interim results are suggesting a lack of efficacy, this too can lead to potential harm if the decision made was incorrect. As previously reported in the background section of this paper, there are examples where a trial was stopped early for reasons of benefit, but follow-up data failed to confirm the interim findings. In the literature, this is most often documented and discussed with respect to closing early on the basis of benefit (i.e. rejecting the null hypothesis). In these cases, there have been instances where the follow-up or long term data (or in evidence from new trials investigating similar hypotheses) have not shown statistical significance for efficacy that was seen at the interim. Incorrectly

closing a trial early, therefore, can also result in potential harm and risk to patients, as decisions regarding the most optimal treatment strategies available for patients are negatively affected. In the case of closing incorrectly for reasons of futility, it can also mean that a sponsor or industry misses out on showing the scientific community that this novel therapy actually does work. It is exactly this argument that has resulted in a difference in opinion in the scientific community for the trend of closing trials early based on interim findings. Some believe that trials should not be stopped early unless there are outstanding concerns for the safety and scientific integrity of the trial, while others feel that there are cases when this should be done. Again, this variation of opinion in the scientific community is documented more when stopping a clinical trial for benefit, but the same principle remains: a dichotomy exists between stopping a clinical trial early for patient interest and safety, and the benefit of the medical community (whether that be cost or resource consumption or quick dissemination of new and practice altering information), versus the jeopardization of accurate scientific results (e.g. overestimation of treatment effect, loss of potentially useful long term knowledge, favourable toxicity profile for an equally effective treatment, etc.). Furthermore, some investigations appear to suggest that there is a trend towards more aggressive guidelines towards early stopping (Friedlin 2009), which means that there is a potential for increasingly more early terminations of clinical trials on the basis of interim findings. It is therefore very important to show that interim analysis methods are accurate with respect to the projected final analysis findings, and that the proper methods are being employed. More research and investigation on this topic is required.

CHAPTER 2 – Assessment by Literature Review:

2.1 Purpose:

The goal of this chapter is to present the results of a systematic review of published phase III randomized clinical trials in oncology that investigated and described the methodological and statistical principles of futility analysis in clinical practice. The conclusions from this chapter serve to inform the principles and methodologies of the next chapter of this investigation (the retrospective futility analysis), where the selected FA methodologies will be evaluated and compared in two NCIC CTG trials.

2.2 Objectives:

Co-Primary Objectives: To describe the nature of randomized phase III clinical trials in oncology that have closed early due to findings of futility by performing a systematic review of RCTs that have included FA, and to inform and apply these findings in chapter 3 of this study: the retrospective futility analysis.

2.3 Study Design and Methods:

In order to investigate the nature of futility analysis, a systematic review of the literature was performed. Eligible articles were selected from the literature and from two other data sources, extracted, and described in detail using a data abstraction key (see Appendix II). The cumulative results of these data were then abstracted and described in detail. A formal data comparison study design is not practical for this type of analysis, since it is methodologies and settings being investigated, rather than a comparison of two or more specific treatments or exposures (investigational entities). In addition to this systematic review, an application of a futility analysis will also be performed by taking the principles learned from this experience, and applying these to two NCIC CTG trials that may have included an interim analysis, but not a futility analysis. This will be described further in Chapter 3.

2.3.1 Eligibility Criteria (Study Selection):

Selected studies include randomized phase III clinical trials in oncology that closed early for reasons of futility. The trial must have included some type of intervention, even if the primary outcome involved secondary benefits (i.e. patient reported outcomes, toxicity, etc., as opposed to overall or performance free survival or response). All such studies identified were included in this analysis.

2.3.2 Ineligibility Criteria:

- Articles not available in English
- Phase I or II studies (including phase II-III 'go-no go' studies)
- Studies not involving an intervention (i.e. studies not involving a patient outcome measure, such as screening or prevention strategy studies)
- Long term follow-up studies or other studies reporting post primary analysis findings (including final analysis studies not reporting IA findings)
- Studies that included reporting from multiple trials
- Articles that did not include an interim analysis (or an interim analysis was not explicitly stated in the body of the article)

2.3.3 Data Sources:

Data for this thesis were collected from two principle sources. The primary data source for the systematic review was derived from a search of both Ovid MEDLINE and EMBASE databases (all databases available at Queens University) to identify trials stopped for futility up until February 2010, when the search was performed. A second source of data for this systematic review was collected from a cohort of previously extracted studies obtained from two prior studies.

Upon completion of the literature search, additional articles for the purposes of the systematic review were then identified from a previous study investigating the evolution randomized clinical trial design, sponsorship, and data analysis in cancer research (Booth 2008). In this previous study, data was extracted from the 6 highest impact journals that publish results of oncology RCTs: *Journal of Clinical Oncology*, *Journal of the National Cancer Institute*, *Cancer Treatment/Chemotherapy Reports*, *New England Journal of Medicine*,

Lancet, and *Journal of the American Medical Association*. The researchers searched these journals for all RCTs of systemic therapy in breast, colorectal, and non-small cell lung cancers published between 1975 - 2004. A total of 321 eligible articles were found for their analysis (48% breast cancer, 24% colorectal cancer (CRC), and 28% lung cancer (NSCLC)), and a hard copy of this entire cohort of articles was made available for the purposes of this study.

After review of the first cohort of previously extracted studies, additional articles were also obtained from another previous study that investigated the nature of RCTs closed on the basis of benefit (Booth 2011). While the focus of this review was to identify trials stopped for reasons of benefit (positive trials), a number of trials stopped for lack of benefit, or futility, were also obtained. The purpose of these two additional data sources was to increase the sample size of the current investigation, and to serve as a validation tool for the literary search results.

2.3.4 Data Extraction and Synthesis:

Literature Search: With the assistance of an experienced reference librarian employed at Queens University, an elaborate data extraction strategy was developed and employed. The literature search was started using MeSH terms 'clinical trial' and 'neoplasm' to identify all randomized clinical trials in oncology. Next, search terms relating to interim analysis, DSMC, and early termination of clinical trials were identified and used to refine the search. These included 'interim analys$', 'futil$', 'exp treatment outcome' (MeSH term), and 'exp methodology' (MeSH term) in the EMBASE directory, and the terms 'interim analys$', 'futil$', 'exp Research Design' (MeSH term), and 'Medical Futility' (MeSH term) in the Ovid MEDLINE database. These search terms were then applied to the pool of 'clinical trials in oncology' for each database. From the final number of 'hits', duplicates and articles not written in English were extracted. The search was performed in February of 2010, and resulted in a total of 628 articles. These articles were exported into RefWorks for proper referencing and organization, and also in order to access the full articles.

Next, each of these 628 articles were reviewed against eligibility/ineligibility criteria, so that only phase III randomized clinical trials including an interim analysis were identified. This pool was then refined by reviewing for all articles that involved closing early at interim analysis for reasons of futility. If there was

uncertainty as to whether or not a trial was closed early for lack of efficacy (futility), the article was reviewed again by Dr. Ralph Meyer in order to obtain a final decision regarding the articles eligibility for this study. The final number of remaining eligible articles extracted from this review was 42, and these were then analyzed for the purposes of this investigation using a data abstraction key (Appendix II).

Cohort of Previously Extracted Studies: The cohort of studies from both additional data sources described above was reviewed, and all journals that had been previously identified as having an interim analysis were extracted. These full text articles were then examined to determine if the reason for early termination was due to findings of lack of efficacy (futility), and that they met all other eligibility criteria. Any additional articles that had not already been identified by the systematic review were included in the analysis. The search from the first cohort of 321 articles resulted in a total of 32 'hits' reporting early trial closure. Only five of these articles met the inclusion criteria and had not been already flagged from the literature search, and these were included in the sample. The second cohort of articles from the study investigating trials stopped for benefit produced another 15 articles that reported closing early for lack of efficacy; all of which had been previously flagged by the literature review. The total sample from all sources to be described, therefore, was 47 articles. For a flow chart of the data extraction process for the systematic review, please refer to Appendix I.

2.3.5 Analysis Strategies:

In order to assess and describe the nature of futility analysis in practice today, a data abstraction key was developed. The data abstraction key can be seen in Appendix II, and was used to extract valuable information pertaining to the nature of the clinical trial itself, its analysis, and its interim analysis methods. The variables extracted from each eligible article using the data abstraction key were used to create a data set in an excel spreadsheet. This compiled data set was then analyzed and described using descriptive statistical methods. The intent of this aspect of the study is to inform the next chapter of this study (i.e. apply the findings of the review to the retrospective futility analysis), so only descriptive statistics were used: proportions were reported for categorical variables, and median and range were reported when applicable. The findings of this analysis are summarized and described in full below. When possible, variables not provided in the actual article were calculated and entered in the spreadsheet, provided that enough data was available to perform these calculations.

Futility analysis methodologies were determined based on descriptions of methods used, or by direct references provided in the article. For determination of aggressive versus conservative stopping guidelines, the five groups of futility boundary as defined by Friedlin and Korn (Friedlin 2009) were used. These boundaries and common futility analysis methodologies are described in the data abstraction key provided in Appendix II.

2.3.6 Potential Confounders:

While overall limitations of this study will be described further in the Discussion section of this paper, it is important to describe some of this information in this section, as the challenges of retrieving futility analysis data from published articles relates directly to this systematic review.

First, it is important to note that futility analyses are essentially a snapshot of the data at a particular (often pre-specified) time point during the course of a trial. In order for recommendations from futility analysis to be considered, one must assume a constant pattern of data to be observed throughout the course of the trial, which may not always be true. Some very conservative trials, therefore, do not set stopping boundaries for futility, or do not consider futility results at the interim altogether, because early data is unstable and it is difficult to assume a uniform distribution of treatment effect. Articles reporting on trials that had developed their futility analysis methods in this manner, therefore, would not have been flagged by the review, and therefore would not contribute information to the analysis of interim analysis parameters and methods.

Another limitation to this review is that often, when trials do not detect any extreme observations at interim analysis, they will not describe their methods or results in any detail in the published article. In order to limit the length of the journal submission, articles are edited considerably, and are not likely to report any non-essential findings. A trial that has been stopped early, however, will likely provide more details of the methodological processes used, in comparison to a trial that proceeded to completion. In other words, if a trial conducted an interim analysis (which may or may not have included an analysis of futility) that did not detect any significant findings, the interim analysis may not be reported adequately in the article. This leads to a type of reporting bias. This bias could potentially result in a number of articles being missed in the literature search, since the interim analysis information is not reported. In fact, this type of reporting bias became apparent quite early in the development and data extraction phase of this systematic review. Originally, the search parameters

included all articles that reported having an interim analysis. Once these articles were flagged, the intention was to review their methodologies, and consider if a futility analysis was included. If a futility analysis was included, and all other eligibility criteria were met, these articles would also have been included in the sample regardless of whether or not the trial closed for reasons of futility. Unfortunately, however, a majority of articles that reported an interim analysis would rarely provide specific details of the interim analysis (i.e. investigations, methodology, findings, etc.), unless there had been significant outcomes to report. Often they would simply mention that there were interim analyses included, but would provide no other details. Thus, the nature of the data to be abstracted was of poor quality, as the missing information was considerable. To address these concerns, the investigation was modified to focus only on articles reporting early closure for lack of efficacy (futility), in order to limit the poor quality data being obtained from articles providing very little information on interim analysis. The intent being that by limiting the search to only articles that reported closing early for reasons of futility based on significance evidence suggesting a lack of efficacy, the reporting bias would be minimized since these articles were most likely to report this information. They would also be more likely to report the interim analysis methodology and design used to obtain these results and justify truncating the trials.

Of course, the most extreme case of bias would be cases where a trial was closed for reasons of futility, and the resulting article describing these findings was not published at all. This is an example of publication bias. While this loss of potential information from both publishing and reporting bias was a concern, it was hoped that this could be minimized by focusing only on articles that reported significant outcomes at interim regarding futility analysis. Additionally, the intent of this review is to describe the nature of trials that closed early for reasons of futility (and their statistical methodologies), and to inform the next chapter of this paper: the retrospective futility analysis. As can be seen from the Results section, a substantial number of articles demonstrating futility were identified using the methods described above, and these articles reported futility analysis methodology and results. As a result, the systematic review was able to effectively inform the next chapter of this study, and therefore, able to achieve the identified goals of this review.

2.4 Ethical Considerations:

Because this aspect of the study only involved data collection and analysis from articles that were already completed and published, and because there were no additional requirements of patients enrolled on these studies, there were very few ethical considerations. The articles were obtained from public records, and the analysis itself was carried out on the reported study methodologies and findings, not the actual patient data. This study, however, was submitted to the Queens Research Ethics Board (Queens HSREB) in order to approve the retrospective futility analysis described in the next chapter, as it involved a secondary analysis to clinical trial datasets from trials that have been previously cleaned and whose results have been published. Documentation of ethics approval is provided in Appendix V.

2.5 Results:

As described above, a total of 47 articles were identified from the literature search and additional datasets for the purposes of this investigation. The data retrieved from these eligible articles are divided into four categories, which will be discussed individually.

2.5.1 Demographic Information:

Table 1 describes the characteristics of the study cohort demographic. The most common disease sites were gastrointestinal (8/47, 17.0%), breast and lung (7/47, 14.9% respectively), and genito-urinary (6/47, 12.8%). The most common setting was metastatic first line therapy trials (28/47, 59.6%), though it should be noted that this group also included limited stage trials. That is, in addition to articles reporting on first line therapy for metastatic disease, trials that included specific advanced stages of disease as part of eligibility (i.e. patients with either metastatic or advanced but possibly resected disease), as well as trials with inoperable disease, were also included in this group. The most common primary endpoints were overall survival (18/47, 38.3%) and disease control (15/47, 31.9%), and the majority were investigating for superiority as opposed to non-inferiority (45/47, 95.7%). The majority of trials involved systemic therapies (37/47, 78.7%), were open trials (35/47, 74.5%), and included an active treatment as the control arm (37/47, 78.7%). In addition, the majority of the articles had been reported in the *Journal of Clinical Oncology* (25/47, 53.2%), and were

published since 2000 (33/47, 70.2%). In fact, only one article meeting inclusion criteria was published prior to 1990 (2.1%), which is consistent with trends of adoption of DSMCs and interim analyses in clinical trial methodology.

Table 1: Demographic Information, Design, and Primary Endpoints of Eligible Articles

Disease Site:	Frequency	Percentage
Breast	7	14.90%
Gastrointestinal	8	17.02%
Lung	7	14.90%
Genito-urinary	6	12.80%
Gynecological	5	10.60%
Melanoma	4	8.50%
Brain	3	6.40%
Symptom Control	2	4.30%
Sarcoma	1	2.10%
Head & Neck	1	2.10%
Hematologic	1	2.10%
Other	2	4.30%
Setting:		
Metastatic: 1st line *(includes limited stage trials)*	28	59.57%
Metastatic: >1st line	3	6.40%
Non-Metastatic: Neo-Adjuvant	3	6.40%
Non-Metastatic: Adjuvant	6	12.80%
Not Applicable *(symptom control or surgery)*	7	14.90%
Primary Endpoint:		
Overall Survival	18	38.30%
Companion Reported Outcome *(Symptom Control)*	6	12.80%
Disease Control - PFS	12	25.50%
Disease Control - EFS	3	6.40%
Response	7	14.90%
Response + Overall Survival	1	2.10%
Control Group:		
No Active Treatment *(placebo/no tx)*	10	21.30%
Active Treatment *(SOC or new tx)*	37	78.70%
Blinding:		
Open	35	74.50%
Single	1	2.10%
Double	11	23.40%
Intervention:		
Radiotherapy	3	6.40%
Systemic Therapy	37	78.70%
Surgery	2	4.30%
Supportive Care/Symptom Control	5	10.60%
Equivalence Study:		
Superiority Trial	45	95.70%
Non-Inferiority Trial	2	4.30%
Collaboration with Industry:		
Yes	10	21.30%
No	37	78.70%
Industry Only	0	0%

Table 1: Demographic Information, Design, and Primary Endpoints of Eligible Articles		
Participants/Sites:		
Multi-centre Trial	42	89.40%
International	13	27.70%

Application to Retrospective Futility Analysis: While the demographic information is helpful to understand the nature of the cohort extracted from the systematic review, the next chapter will focus primarily on the futility analysis methods that were identified. The trials selected for performing retrospective futility analyses, however, will include a gastrointestinal (GI) and a gynecological (GU) trial that both investigated systemic therapies in a metastatic setting. Both are open label and multi-centered international trials that attempted to demonstrate superiority of the treatment arm versus another active treatment (standard of care). The primary endpoints for these trials were overall survival and progression free survival, and both trials will be described further in chapter 3.

2.5.2 Design Statistics and Results:

Table 2 describes the overall design statistics and results for the identified articles. The median expected hazard ratio was 0.67 (range 0.4, 2.73), and the majority of trials were looking for a hazard ratio of about 0.6 to 0.79 (19/47, 40.43%). Thirteen of the articles were reported as 'not applicable' for expected hazard ratio, because they were either response or symptom control trials, and an expected hazard ratio is generally not provided for these types of trials. Five articles in the final cohort with a primary endpoint as either survival or event free survival, however, did not report an expected hazard ratio, and did not provide adequate information in order to calculate this statistic. These were listed as 'unknown' and account for10.6% of the sample.

The majority of articles provided an expected absolute benefit in percentage, but 5 of the 47 articles used 'months' as the unit of measure for this variable (10.6%). The median for expected absolute benefit was 20% (range 7, 50), and 12 months (range 3, 20) for those articles that reported this value in months instead of as a percentage. Twelve of the 47 articles did not provide an expected absolute benefit (25.5%), and of these twelve articles, 7 did not report either an expected hazard ratio or absolute benefit value (14.9%).

Samples sizes for the cohort varied significantly. Median expected sample size was 293 (range 60, 1800) with 4 trials not providing this information, while the median of the actual sample size obtained was

calculated as 167 (range 39, 4312). The article that reported a final sample size of 4312 (high end of the range), unfortunately, did not provide an expected sample size value. A percentage of sample size obtained was also calculated for the articles that provided adequate information, and the median for this value was calculated as 62.4% (range 28.9, 100).

Planned power was most frequently in the 80 – 85% range (28/47, 59.6%), while the planned two-sided alpha was most frequently 0.05 (27/47, 57.4%). Power was not reported in 7 cases, and alpha was not reported in 6 cases (5 of these 7 cases did not report either a planned power or alpha). An actual final p-value was provided in the majority of cases (42/47, 89.4%). Of the five articles that did not report a final p-value, two also did not report a hazard ratio or relative risk, and did not provide enough information for these to be calculated. Twelve articles reported inferiority of the treatment arm (p-value was found to be significant, but in favour of the control arm), and 15 of the 47 articles had a p-value that was greater than 0.7 (31.9%). Again, all articles reported that the trial had been stopped for reasons of futility.

Table 2 - Design Statistics and Results

Expected Hazard Ratio:	Frequency	Percentage
0.4 - 0.59	7	14.90%
0.6 - 0.79	20	42.55%
0.8 - 0.99	0	0.00%
>1 (non inferiority)	2	4.26%
Unknown	5	10.60%
Not Applicable	13	27.70%
Expected Absolute Benefit:		
0 - 9.9 (%)	1	2.10%
10 - 19.9 (%)	10	21.30%
20 - 29.9 (%)	4	8.50%
30 - 40 (%)	10	21.30%
>40 (%)	5	10.60%
0 - 10 (months)	2	4.25%
11 - 20 (months)	3	6.40%
Unknown	12	25.50%
Power (Planned):		
<80	1	2.10%
80 - 85	28	59.60%
86 - 90	8	17.00%
>90	3	6.40%
Unknown	7	14.90%
Alpha (Planned):		
0.1	10	21.30%
0.05	27	57.40%
0.025	3	6.40%
0.2	1	2.10%
Unknown	6	12.80%

Table 2 - Design Statistics and Results		
Alpha (Obtained):		
Inferiority (significance)	12	25.50%
<0.2	4	8.50%
0.2 - 0.5	9	19.10%
0.51 - 0.7	2	4.30%
0.71 - 0.9	11	23.40%
>0.9	4	8.50%
Unknown	5	10.60%
Sample Size:	Median	Range
Sample Size (obtained)	167	(39, 4312)
% of Sample Size Obtained	62.40%	(28.9, 100)

Application to Retrospective Futility Analysis: Both trials to be used in the retrospective futility analysis described in the next chapter included a two sided significance level of 0.05, and a power of 80% or higher (90% for CO.17). One of the trials resulted in significant findings in favour of the treatment arm, while the other did not detect a significant difference between treatment arms.

2.5.3 Timelines for Trial Completion:

Table 3 summarizes the timelines reported in the identified articles. An interesting trend in the data extracted, however, was the number of articles that did not provide the expected timelines for trial completion and accrual. Only 13 of the 47 articles reported both timelines for reaching target accrual, and timelines for follow-up (i.e. timelines for reaching primary analysis). This amounted to only 27.7% of the sample which provided this important information, and 59.6% did not report any expected timeline information at all. Reporting of observed timelines for accrual and follow up, however, was considerably better, with 21 of 47 articles providing this information (44.7%). Despite this, there was still considerable data missing, as 17% did not report the timelines at which the trial was truncated for lack of efficacy. In addition to these numbers, only 31.9% of the articles provided an expected accrual rate (or one could be calculated from the data provided), while the actual accrual rate could be obtained in 78.7% of the articles identified. This made it extremely difficult to provide adequate numbers for ‘percentage of expected accrual rate actually obtained’, and resulted in 72.3% of the cases being captured as unknown due to lack of information. As previously mentioned, expected and actual accrual values were well reported, and this allowed for a calculation of percentage of expected accrual obtained at the time of trial closure. The results of these calculations varied considerably throughout the

data. 42.6% of these trials, however, had only accrued 59% or less of their expected accrual. Overall, reporting of timelines and expected accrual rates were poorly reported in these articles, which may be a trend in RCT reporting and publications.

Table 3 – Timelines for Trial Completion		
Timelines for accrual and follow-up Reported? (expected):		
Both Reported	13	27.70%
Accrual Only	1	2.10%
Follow up Only	5	10.60%
None Reported	28	59.60%
Timelines for accrual and follow-up Reported? (observed):		
Both Reported	21	44.70%
Accrual Only	14	29.80%
Follow up Only	4	8.50%
None Reported	8	17.00%
Percent of Accrual Obtained:		
100 or more	5	10.60%
90-99	4	8.50%
80-89	2	4.30%
70-79	6	12.80%
60-69	6	12.80%
50-59	11	23.40%
<50	9	19.10%
Unknown	4	8.50%
Accrual Rate:		
Accrual Rate – Expected (provided or calculated)	15	31.90%
Accrual Rate – Observed (provided or calculated)	37	78.70%
Percent of Accrual Rate Obtained:		
100 or more	2	4.30%
75-99	2	4.30%
50-74	4	8.50%
25-49	5	10.60%
<25	0	0.00%
Unknown	34	72.30%

2.5.4 Methodology of Futility Analysis:

A summary of the findings obtained from this review can be seen in Table 4. Twenty nine of the 47 articles explicitly provided a pre-defined statement of futility (61.7%), including articles which simply indicated which methodology they used without any further details. Four of the identified articles indicated that the interim analysis, which included an analysis of futility, was unplanned (8.5%). In these cases, the interim

analysis was prompted by concerns for the trial expressed by either investigators or DSMC. These included concerns for accrual, concerns for safety, and concerns for futility. Interestingly, 13 of the 47 articles did not report the methodology used for the futility analysis that ultimately provided the evidence to stop the trial early (27.7%). Of these 13 articles, three also did not report a final p-value or hazard ratio (6.4%).

The two most frequently used methodologies were testing the alternative hypothesis at very low significance level (as originally proposed by Fleming and O'Brien (Fleming 1979)), and stochastic curtailment methods (Ware 1985) at 25.5% and 12.8% respectively. These methodologies will be described further in the next chapter of the study. Three articles did not report a primary methodology for futility analysis, but did report multiple methodologies used in the interim analysis. These included a combination of testing the alternative hypothesis at low significance level (2/3), stochastic curtailment (2/3), and error spending (1/3) methods, and these three articles were reported separately in Table 4. Five of the six articles that reported a primary futility methodology as stochastic curtailment reported a planned conditional probability statistic. Interestingly, although only eight articles in total included conditional probability as part of their pre-defined statement of futility analysis, a total of 15 articles (31.9%) reported a conditional probability to support the rationale for truncating the trial. Conditional Probability could also be calculated for an additional 24 articles as well, for a total of 39 of 47 (83%). The majority of these articles reported a conditional probability of 2% or less (29/39, 74.4%), and all of them reported a conditional probability of less than 10% (highest was 8.6%). Forty-one of the articles reported 3 planned interim analyses or less (87.2%), with only one interim analysis being planned in 27 cases (57.4%). Criteria used for interim analysis was most often reported number of patients accrued (24/47, 51.1%), while 10 articles did not provide the criteria to trigger an interim analysis (21.3%). Upon reviewing the data accumulated and reported at the time the trials were stopped, it was determined that 26 of the 47 articles were conservative when deciding to close the trial (55.3%), while 23.4% were more aggressive in reaching a decision to stop the trial. These conclusions were based on the five groups of futility boundary proposed by Friedlin and Korn (Friedlin 2009). Ten of trials did not provide adequate information to determine whether the futility methodology was conservative or aggressive (21.3%).

Table 4 - Futility Analysis Methods			
Pre-Defined Statement of Futility:		Frequency	Percentage
Yes		29	61.70%
No		14	29.80%
Unplanned Analysis		4	8.50%
Futility Methodology:			
1 - Low Significance Level		12	25.50%
2 - Error Spending		1	2.10%
3 - Power Family (wedge tests)		0	0.00%
4 - Stochastic Curtailment		6	12.80%
5 - Whitehead Triangular Test		3	6.40%
6 - Bayesian Methods		3	6.40%
7 - Pre-Specified Stopping Boundary		4	8.50%
8 - Multiple		3	6.40%
Unknown		13	27.70%
Conditional Power:			
Was a conditional Power reported?		15	31.90%
Could one be obtained?		24	51.10%
Total Conditional Power values obtained:		38	80.85%
Range of conditional power (%):			
0 – 1%		25	65.79%
1 – 5%		9	23.68%
5 – 10%		4	10.53%
>10%		0	0%
Number of Interim Analyses (Planned):			
1		27	57.40%
2		7	14.90%
3		7	14.90%
>3		2	4.30%
Unknown		4	8.50%
Criteria Used for Interim Analysis:			
Patients		24	51.10%
Events		12	25.50%
Both		1	2.10%
Unknown		10	21.30%
Decision to Stop:			
Moderately Aggressive Early Stopping	(score = 1)	7	14.90%
Aggressive Early Stopping	(score = 2)	2	4.30%
Moderately Aggressive Late Stopping	(score = 3)	2	4.30%
Aggressive Late Stopping	(score = 4)	0	0.00%
Conservative	(score = 5)	26	55.30%
Unknown		10	21.30%

Application to Retrospective Futility Analysis: The two most common futility analysis methods were the 'testing the alternative hypothesis at a very low significance level' (according to Fleming and O'Brien), and the

'stochastic curtailment method based on conditional power (futility index)'. These will be the two methods included in the retrospective futility analysis. These methods will be tested in both aggressive and conservative settings, since there was variability in aggressive and conservative stopping rules used in the literature, and the FA methods will be investigated at three separate time points in each of the selected NCIC CTG clinical trials. This means that both methods will be compared in three pre-specified retrospective interim analyses for each of the two NCIC CTG clinical trials, and using both conservative and aggressive stopping rules.

2.6 Conclusions of Review

In summary, the overall objective of this study is to improve understandings of design, conduct and analysis of randomized clinical trials by evaluating the methodological and statistical principles associated with decisions to close trials early at interim analysis due to evidence of futility. The first step in achieving this goal was to perform a structured systematic review designed to identify and process eligible articles of phase III randomized clinical trials in oncology. A total of 47 eligible articles were identified from numerous sources, data was extracted from these articles using a data abstraction key, and analyzed using descriptive statistics. The specific co-primary objectives were to describe the nature of these eligible articles in order to increase knowledge of futility analysis methodology currently used in practice, and then to apply these results to the second phase of this study: the retrospective futility analysis. While this systematic review was able to meet its co-primary objectives by obtaining a substantial amount of information as to how futility analysis is performed in practice in order, more research is required in order to adequately address the study's ultimate objective. The next phase of this study, therefore, will be designed as an application of the major findings of this review, and will aim to retrospectively perform futility analyses (based on the information obtained from this review) in two previously published NCIC CTG phase III clinical trials in oncology. It will then compare these retrospective interim findings with the actual final results of the original data set from these trials in an effort to provide more insight into the effectiveness and accuracy of each futility analysis method, and to provide recommendations for interim futility analysis planning when designing future RCTs in phase III oncology.

Although this review was able to provide a considerable amount of information, it was interesting to see the abundance of unknown values obtained from the literature. Despite attempting to minimize the contribution of articles providing poor quality data (i.e. articles that provided very little interim analysis information) by focusing on articles that reported closing early for lack of efficacy, a trend towards underreporting interim and futility analysis methodologies and results was still evident in the compiled data. This observation was not expected, as it was assumed that if interim findings were significant enough to merit closing a trial early on the basis of futility assessments, then the analysis plan that led to these conclusions would also be provided. This was often not the case, or at least not all the required data was available or reported. Information regarding study demographics was well reported as can be seen from Table 1, but missing data regarding original design statistics from the articles is still observed in the review. Seven of the 47 articles did not report either an expected Hazard Ratio or expected absolute benefit (14.9%), and over 10% of the articles did not report the desired power or alpha statistics (5/47, 10.6%). Four articles also did not report their sample size target. All of the eligible articles in the review, however, reported the actual sample size obtained during the course of the study. Another 10.6% of the articles (5/47) also did not report the final p-value obtained, and two of the articles did not report a final p-value, futility analysis methodology, or other final results to justify closing the trial early. These articles reported stopping early for lack of efficacy, but did not explicitly provide the results to justify the conclusion (though this information was implied when describing their findings). In fact, over 25% of the articles did not explicitly state the futility analysis methodologies used to reach the conclusion to stop the trial (13/47, 27.7%), over 10% did not provide enough information to calculate a conditional probability (8/47, 17.0%), and over 20% did not provide adequate information to determine if they used aggressive or conservative futility methods (10/47, 21.3%). A number of cases also did not report the criteria used for interim analysis, although this could be implied depending on the type of trial reported in the article, and a few articles also did not report the number of interim analyses included (4/47, 8.5%). This underreporting of futility analysis methods was concerning and a very interesting finding given the focus of this study.

Although arguably more of an issue with trial feasibility and accrual, another very poorly reported area was the timelines for the trials. Almost 60% of the articles did not report any timelines for accrual or trial completion (follow-up) at all, and less than 30% reported both of these important design parameters. As a result,

the calculation of accrual rates was lacking because of this missing information, and a percentage of expected versus actual accrual rates obtained could not be calculated in over 70% of the cases. This is unfortunate as these values would suggest how feasible the trials were when they closed, and also could have potentially provided insight into areas in oncology that may require a focus on increasing accrual strategies. This also was a barrier for determination of aggressive versus conservative stopping in these articles.

Focusing on the positive, however, a number of important findings were observed from the data obtained in this review. This information was able to adequately inform the retrospective futility analysis, and provides insight on how futility analyses are planned and designed in a phase III randomized clinical trials in an oncology setting. About 70% of the articles did include a predefined statement of futility, or indicated that a pre-defined statement was not included because an unplanned analysis was undertaken due to some concerning trends in the study. The most predominant futility analysis methodology in the literature was the 'testing the alternative hypothesis at very low significance level' method according to Fleming and O'Brien. This method was reported in 29.8% of the sample, if articles reporting the use of multiple methods are also included (14/47, 29.8%). The second most frequently used methodology was the 'stochastic curtailment method based on conditional power (futility index)', which was reported in a total of 17% of the sample including the articles that reported using multiple methodologies. An interesting concept discovered from this review is that a number of articles that did not use the stochastic curtailment method as the primary methodology for futility analysis included a conditional power statistic as additional support for early closure. While 8 trials indicated stochastic curtailment as a primary methodology, an additional 7 trials included a conditional power value as supportive evidence for early closure (15/47, 31.9%). The conditional power could also be calculated in an additional 24 articles, and in almost 75% of the articles reporting early closure for futility, this statistic was calculated as being less than 2% (overwhelming evidence for closure). A conditional power of less than 10% was obtained in all of the articles where this statistic could be determined, which is the cut off proposed by Friedlin and Korn (Friedlin 2002, Friedlin 2009) regarding early stopping rules in favour of the null hypothesis (assuming 50% of data was obtained). Because of the prevalence of these two futility analysis methods in the literature, these two methodologies will be applied and compared in the retrospective futility analysis. This was the first major finding from the review.

A number of other futility methods were identified in this investigation as well, including Bayesian methods and simply pre-specifying a stopping boundary. For example, some articles reported that if the hazard ratio was found to be >1 at any time during the interim analyses, then the trial should stop. Another reported that if there were no complete or partial responses observed in the treatment arm by x amount of patients, then the trial should stop. A total of 3 articles also referenced Bayesian methods for interim analysis, including one that referenced specific Bayesian methodology software called Mathematica.

The majority of trials suggested conservative approaches for their futility analysis results, and subsequent decisions to close the trial early. That is, with 50% or more of the required information, the trial should continue unless the hazard ratio of the new over the standard is greater than 1. This trend seems reasonable, since deciding to close a trial without all the data available may introduce a chance of error, and the consequences of incorrectly advising the discontinuation of a trial can have an impact a number of stakeholders including patients, sponsors, industry, and investigators. As previously mentioned, there are potentially very serious consequences of incorrectly closing a trial early, such as decreasing the potential risk subjected to patients or possibly for financial reasons and resource management. It is important to note, therefore, that the majority of the articles identified in this report did so with overwhelming evidence to support the decision. Roughly 15% of the articles, however, did report either moderate aggressive (7/47, 14.9%), or more aggressive (2/47, 4.3%) early stopping. These trials would have been stopped with less than 50% of the information available, if they could reject the alternative hypothesis at the 0.001 or 0.005 level respectively. Given the variation in the results between conservative (55.3%) and aggressive stopping methods (23.4% total), it would be informative to incorporate these guidelines in the retrospective futility analyses of this study, and compare these settings to the final analysis findings. A comparison of these stopping rules, therefore, will be included in the next chapter of this study. The variability in the use of conservative versus aggressive stopping rules was the second major finding from the review.

In conclusion, there appear to be some trends of futility analysis identified from this review. These trends have provided important insight into how futility analyses are executed in the literature (and in practice), and will be applied to the retrospective futility analysis in order to further investigate their accuracy and benefit. Unfortunately, however, a significant trend observed in this review was the systematic underreporting of these

futility methods in published papers, even in a setting where the results of these interim analyses are substantial. The lack of reporting of expected versus observed timelines is also noted in this sample as well, and this seems to be a common trend in articles reporting on randomized phase III clinical trials. These findings are likely due to constraints for articles submissions. Quite simply, the full reports of clinical trials are often edited considerably for publication, and therefore information considered less 'important' is omitted.

It could be argued as well that the findings from this review also displayed the divide in scientific opinion for how best to perform futility analysis in RCTs. When futility analysis methods were reported, there was considerable variation in the methodologies used, as well as variation in the use of conservative and aggressive stopping guidelines for deciding to close a trial.

Despite the barriers experienced during this review, this aspect of the study was a very useful exercise in investigating futility methods used in randomized phase III clinical trials in oncology. Two frequently used futility methods were identified from the literature, as well as considerable variation in the use of conservative versus aggressive guidelines for stopping clinical trials (and futility analysis methodology in general). These important findings can be taken forward and applied in the next chapter of this study, and, therefore, this systematic review was successful at achieving the goals outlined at the onset of the study. In these economic times, there is a focus on the bottom dollar costs for clinical trials, how best to ensure that the limited resources available are being applied in a most effective method, and in a way that exposes patients to the least amount of foreseeable risk as possible. Large scale (phase III) randomized clinical trials are very expensive and complicated endeavors. Their designs are becoming more complex, as are the agents that are being investigated. These large scale clinical trials are also becoming more of a global coordination as new international collaborations are executed, which increases both complexity and costs. In this business savvy, expensive, and complex environment it is very important to have appropriate and consistent measures in place to continually test the relevance of these large RCTs, so as to avoid risk to both sponsor (whether industry, shareholders, cooperative group, or even local investigator and hospital) and patient. Systematic reviews such as this one, therefore, are an important investigation of appropriate methods for designing and applying futility analyses so that inefficacious trials may be stopped early in a manner that is accurate, appropriate, and safe. This study was able to identify some important trends with respect to these futility analysis methods. As such, it will be

beneficial to investigate these trends further in the second phase of this study, in order to gain further insight into the effectiveness and accuracy of these methodologies, and in order to achieve the overall objectives of the study. In addition, because of the variability in the reporting quality of the data extracted from the literature, a recommendation for future research is to provide more information regarding how decisions about trial closure are made when reporting/publishing these findings. Inclusion of this information would be useful for future research and reviews of interim and futility analysis methodologies, and would serve as guidance for other researchers wanting to incorporate these methods in their protocol development. When an RCT closes early for findings of lack of efficacy during the course of a trial, it is important to understand and report the methodologies and findings used to arrive at these important decisions. The consequences of closing a trial early can be quite substantial for a number of stakeholders, as well as for future or even concurrent research projects.

CHAPTER 3 – Application of Futility Analysis Methods:

3.1 Purpose:

The goal of this chapter is to report an evaluation of the methodological and statistical principles associated with the conducting analyses of futility, by retrospectively applying the results of the Systematic Review in two NCIC CTG led clinical trials that have previously published primary final analysis findings. Neither of these two trials included interim futility analyses in their original designs. The overall objective of the project is to improve understanding of design, conduct and analysis of randomized phase III controlled clinical trials in oncology. The hope is that the results may be used to provide insight into the effectiveness and accuracy of futility analysis methods, to promote consistency in clinical practice with interim analysis planning, and to provide a basis for hypotheses-testing of optimum methodologies and their associated trade-offs.

3.2 Objectives:

Primary Objective: To assess the accuracy of frequently used futility analysis methods in both conservative and aggressive settings, by applying these methods retrospectively in two previously completed NCIC CTG led randomized phase III clinical trials in oncology and evaluating the results.

Secondary Objective: To consider if one futility analysis method may yield different results for predicting trial outcome, by exploring the results of these methods at each interim analysis with the results of the final analysis.

Secondary Objective: To assess the accuracy of futility results at the different interim analysis time points in conservative versus aggressive settings, in order to consider if may be more appropriate to use conservative or more aggressive methods when deciding to stop a trial.

3.3 Hypothesis:

The retrospective futility analysis results for the OV16 trial, a trial that did not show a statistical significant difference between treatment and control arms, will demonstrate that the trial could have been stopped early using either futility analysis method. For the retrospective analysis on the CO17 trial, however,

these same methods will not provide sufficient evidence to conclude the trial could have been stopped early. Furthermore, the results of the futility analyses will not differ significantly in conservative versus aggressive stopping settings, and the results will not differ significantly between each method in any instance (i.e. both methods will arrive at the same result each time).

3.4 Study Design and Methods:

In order to demonstrate the accuracy of futility analysis, the two most common futility analysis methods identified in the Systematic Review will be applied retrospectively to the two NCIC CTG trials that had previously published final analysis results. The two methods will be applied at specific time points in the trials, based on a percentage of the number of events required for final analysis. In order to increase the robustness of the futility methods, 1000 bootstrap samples will be used for each retrospective analysis. That is, the retrospective analyses will be run repeatedly at each time point, for each trial, and for each futility method. This will be done in each bootstrap sample, and a frequency chart will be created to quantify the results. At each time point, both futility analysis methods will be applied using aggressive and conservative guidelines for early stopping. The analyses will provide a distribution of results that can be compared to meet each objective.

3.4.1 Description of Data Sources:

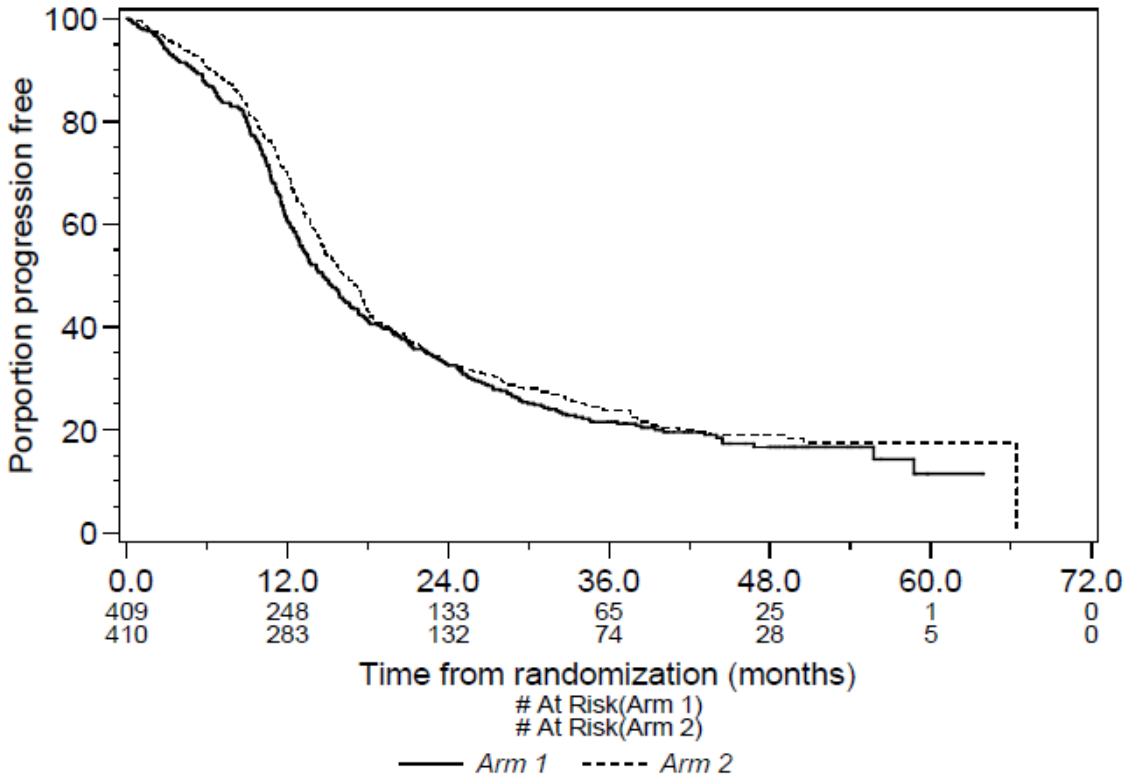
A Phase III Study of Cisplatin Plus Topotecan Followed by Paciltaxel plus Carboplatin versus Paclitaxel plus Carboplatin as First Line Chemotherapy in Women with Newly Diagnosed Advanced Epithelial Ovarian Cancer – OV.16:

This first NCIC CTG led randomized phase III clinical trial was selected because its final analysis did not demonstrate a significant difference between the two arms of the study (Hoskins 2010). In fact, the two arms of study were very similar in efficacy, and the resulting hazard ratio was very close to 1. This trial was selected, therefore, because a futility analysis may have been able to conclude that the trial should have been stopped early due to lack of efficacy of the treatment arm.

The OV.16 trial was designed to compare the time to progression between patients receiving 8 cycles of standard carboplatin/paclitaxel to those receiving 4 cycles of cisplatin/topotecan followed by 4 cycles of

carboplatin/paclitaxel. 800 patients and 631 events (progression) were required in order to demonstrate a 25% relative improvement in progression free survival (i.e. from 16 to 20 months with a power of 80% and two sided alpha of 0.05, HR = 1.25). During the course of the trial, 819 women with stage IIB ovarian cancer (or worse) were enrolled, and monitored for progression and survival information, as well as adverse effects, quality of life and CA125 normalization. The study opened in August of 2001 and closed to accrual in June 2005. The data base was locked in March of 2008, once the required number of events had been achieved, and the data was analyzed on an intent to treat basis. After a median follow up of 43 months and 650 observed progressions, there was no statistically significant difference observed between the two arms of the trial. The median PFS was 14.6 months versus 16.2 months in arms 1 and 2 respectively (HR = 1.10, 95% CI = 0.95 – 1.30, P = 0.25). While the experimental arm did not showed improved efficacy, and therefore the null hypothesis was not able to be rejected, it did show an increase in toxicity, and so treatment of these patients with Carboplatin with Paclitaxel remained the standard of care.

Figure 1: OV.16 – Progression Free Survival



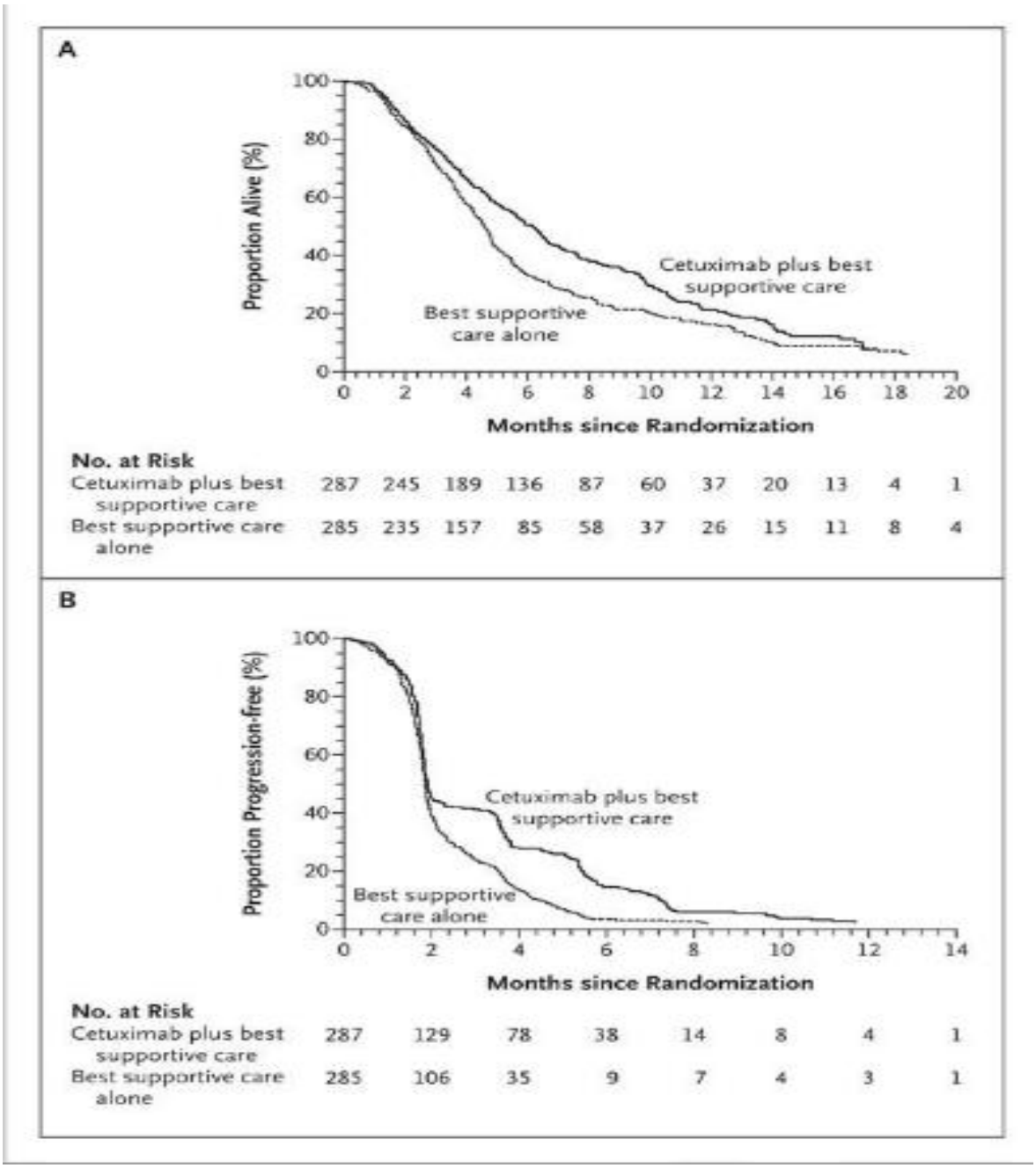
A Phase III Randomized Study of Cetuximab (Erbitux TM, C225) and Best Supportive Care versus Best Supportive Care in Patients with Pretreated Metastatic Epidermal Growth Factor Receptor (EGFR) - Positive Colorectal Carcinoma – CO.17:

This second NCIC CTG led randomized phase III clinical trial was selected because, unlike OV.16, its final analysis did in fact demonstrate a significant advantage in favour of the treatment arm (Jonker 2007). While the advantage was found to be statistically significant, the actual (absolute) differences between the arms when comparing overall survival were not extreme. This trial, therefore, was selected because although it did demonstrate a significant advantage in favour of the treatment arm, the actual difference between arms was marginal. A futility analysis in this case may or may not have concluded that this trial could have been terminated early, as it is unknown how an interim analysis would evaluate these marginal findings during the course of the trial.

This trial was designed to evaluate the effect of cetuximab on the overall survival in patients with advanced, pre-treated (and EGFR positive) colorectal cancer in comparison to best supportive care alone. 500 patients and 445 events (deaths) were needed in order to show a 9.6% absolute improvement of overall survival (power of 90%, two sided alpha of 0.05, HR = 1.36). From November 2003 to August 2005, 572 patients were enrolled on the trial and randomized to receive either cetuximab (400mg/m² loading dose followed by weekly 250mg/m² infusion) and best supportive care versus best supportive care alone. The study met the pre-determined number of events necessary for final analysis in March 2006. The results of the final analysis indicated that the cetuximab treatment was associated with a significant improvement in overall survival, with median survival of 6.1 versus 4.6 months in the cetuximab and control arms respectively (HR = 0.77, 95% CI = 0.64 – 0.92, P = 0.005). This strong signal of effect was a very important finding in this population of metastatic colorectal cancer patients, even though the actual difference in survival was relatively marginal. These findings facilitated the ability for secondary investigations of effect in different populations, as well as subset analyses within the sample to further investigate this treatment effect. An example of this can be seen with the confirmation of poorer outcomes with cetuximab in patients with tumours expressing mutations of a specific protein called KRAS (a subset analysis), which led to additional investigations involving these patients (Karapetis 2008, De Roock 2010).

A total of 456 deaths had been observed at the time of the final analysis, and median follow-up was 14.6 months. The trial also reported that there were similar advantages for the cetuximab arm for progression free survival, preservation of quality of life, and deterioration in physical function and global health status scores. The cetuximab arm, however, was also associated with higher incidence of grade 3 adverse events or higher.

Figure 2: CO.17 – Overall Survival and Progression Free Survival



3.4.2 Key Variables – Futility Analysis Methods:

Testing the Alternative Hypothesis at a very low Significance Level:

This method of testing for benefit and/or futility based on interim results, as proposed by Fleming and O'Brien (Fleming 1984), is a very appealing and popular one or two sided approach. It was first proposed in 1979 (Fleming 1979), and is still widely used today, as can be evidenced from the Systematic Review. It is an appealing type of group sequential approach because it is designed to allow the significance level at the final analysis to be near the desired overall significance. It does this by allowing the interim analysis significance levels to be very small, but increasing such that the sum of all interim significance will equal the desired final significance. For example, final Type I error $\alpha = \alpha_1 + \alpha_2 + \alpha_3 \ldots + \alpha_k$ (k = total number of interim analyses) when investigating rejecting the null hypothesis (benefit). The increase in the size of the interim significance is proportional to the size of the sample size. This is done in order to minimize the potential for making type I errors when investigating for benefit, and type II errors when investigating for futility. Recall that the more times you look at data, the more likely you are to find a significant result based on chance alone. The advantage of this method is that it is difficult to close a trial early when the sample size is still small, and the design is flexible and can be modified to tailor to the trial characteristics. For example, significance levels can be set very low for initial analyses if you want to ensure long term data, or you can set levels very high if you expect large treatment differences. A disadvantage of this method is that the number of interim analyses must be pre-specified, and equally spaced. However, there have been variations of the Fleming O'Brien test that have been introduced to overcome a few of these drawbacks, such as methods to increase k in the event of very slow accrual, or alpha spending functions such as those proposed by Lan and Demets (Lan 1983).

In general, this method is designed to stop a trial early at interim analysis if the test statistic (Z), at a specific analysis (k), is outside of the range of two critical values (a, b) (Pampallona 1994):

- $Z \geq a$ then stop and reject the null hypothesis (stop for benefit) at analysis k
- $Z \leq b$ then stop and accept the null hypothesis (i.e. stop for inferiority or futility) at analysis k

And :

- $a = C_1 (\alpha, ß, K, \Delta) j^{\Delta}$ - where the constant C_1 will depend on the required significance level of the trial (α), the power of the trial (ß), the number of 'looks' or analyses (K), and a parameter Δ, which

is considered = 0 for Fleming and O'Brien methodology. 'j' is equivalent to the particular analysis number.

- $b = j_{\delta} - C_2(\alpha, ß, K, \Delta)\, j^{\Delta}$ where the constant C_2 depends on the same parameters as previously described, and δ is the treatment difference of interest.

The test statistic calculated for this method is a p-value. With respect specifically to futility analysis, however, this p-value is applied to the alternative hypothesis instead of the null hypothesis as is seen with tests for benefit or in final analyses. In this respect, this method would conclude that a trial should be stopped if that p-value were less than a pre-determined boundary. That is, if the calculated futility analysis p-value is less than a specific boundary, one would conclude that there is a very small statistical probability of rejecting the alternative hypothesis, when it is in fact true (i.e. small chance of a type II error). This will be the first futility analysis method employed to both NCIC CTG clinical trials at pre-specified time points during the course of the trials (retrospectively). Critical values will be obtained for each retrospective futility analysis, in both aggressive and conservative settings.

Stochastic Curtailment method based on Conditional Power (Futility Index):

The stochastic curtailment method of futility analysis, as proposed by Ware, Muller and Braunwald (Ware 1985), is a very different approach as it relies on predictive inferences instead of focusing only on currently available data. It involves the calculation of a conditional power statistic, and estimates the likeliness of seeing a positive/negative outcome of a trial based on current information. Instead of setting a specific boundary or threshold at a specific time point (and using that boundary to determine if the trial should close or continue), it attempts to determine if the results of the interim analysis are ever likely to change (i.e. will there ever be a chance that there could be a significant difference). In other words, the conditional probability statistic estimates the probability of being able to reject the null hypothesis at the time of final analysis. If the quantity of this statistic is very small, then one can conclude that it is futile to continue with the investigation. An advantage of this method is that it is more adaptive to trial design, because it does not have an impact on the final p-value of the trial. A disadvantage, however, is that it relies on certain assumptions, such as consistent treatment effect and design effects (Snapinn 2006). As per Friedlin and Korn (Friedlin 2009), a conservative result that would suggest closing a trial early based on reasons of futility would be a conditional probability of less than 0.1 (or

10%). Emerson, Kittelson, and Gillen (Emerson 2005) confirm this proposal, but also suggest a commonly used more aggressive boundary for conditional probability is 0.2 (or 20%). This will be the second futility analysis method applied to both NCIC CTG clinical trials, at the same time points as the first method. It too will be used in both aggressive and conservative settings for each retrospective futility analysis, as described above.

NCIC CTG Trial Data Points:

From each study, a list of variables will be required in order to effectively perform the retrospective futility analyses, and to compare these analyses with each trial's original final analysis results. These variables include overall survival, progression free survival, arms of study (treatment vs. standard of care), stratification factors, sample size, number of events, and analysis statistics (expected and observed absolute benefit, expected and observed hazard ratio, significance values, power, and confidence intervals). The actual obtained final analysis statistics of each trial that were reported and published (i.e. hazard ratio, confidence intervals and p-value, as well as number of events and sample size) will be replicated in order to verify the results. This will be accomplished by repeating (replicating) the original final analysis on the obtained data set of each trial in order to ensure the numbers are accurate, thereby validating our methods.

3.4.3 Time points for Retrospective Futility Analyses:

For both trials, retrospective futility analysis will be performed at 25%, 50%, and 75% of the required number of events for final analysis. The primary endpoint for OV.16 was progression free survival (PFS). In this trial, a total of 631 progression events were required for the final analysis in order to have an 80% power to detect a 25% improvement in PFS using a two-sided 5% alpha (hazard ratio 0.8). The trial also estimated 400 patients per arm (800 total) to achieve this number of events. This means that futility analysis will be performed when 158, 316, and 473 events (progressions) had been observed.

CO.17 required 445 deaths to have a 90% power to detect a 9.6% absolute increase in 1 year overall survival, using a two-sided 5% alpha (hazard ratio 0.74). The primary endpoint for this trial was overall survival. This means that futility analysis for this trial will be performed when 111, 223, and334 events (deaths) had been observed.

3.4.4 Aggressive versus Conservative guidelines for stopping early:

A retrospective futility analysis (using both futility analysis methods) will be performed for each NCIC CTG trial at the time points when 25%, 50%, and 75% of the expected events have been obtained. In addition, each analysis will be done in an aggressive and conservative setting for stopping rules. For aggressive early stopping (analysis at 25% of expected events), a decision to stop the trial due to lack of efficacy will be determined if the alternative hypothesis can be rejected at 0.005 level (i.e. a futility analysis p-value of <0.005). For the interim futility analyses at 50% and 75% of expected events, aggressive rules for stopping will be applied if the alternative hypothesis can be rejected at the 0.02 level (i.e. a futility analysis p-value of <0.02). The conservative stopping rule will involve a decision to stop the trial if the lower confidence interval boundary for hazard ratio of the new/experimental treatment over the standard treatment is >1 (i.e. the experimental arm is worse than control) at the 50% or 75% time points. While the same conservative rule will be applied at the 25% mark, this is purely hypothetical as conservative stopping does not consider a decision to stop/continue a trial when 25% of events have been obtained, as not enough data has been accumulated. These decisions for aggressive versus conservative stopping were originally described by Friedlin and Korn (Friedlin 2009), and these same definitions will be used for the futility method of 'testing the alternative hypothesis at a low significance level'. As previously stated, conservative stopping rules for stochastic curtailment (conditional power) involve a decision to stop the trial if the conditional power is less than 10% as per Friedlin and Korn (Friedlin 2009), and aggressive stopping rules involve a decision to stop the trial if the conditional power is less than 20% as per Emerson et al (Emerson 2005). These stopping rules will be applied at each time point (25%, 50% and 75%), and compared to the results of futility for 'testing the alternative hypothesis at a low significance level'.

With respect to futility analysis stopping boundaries using the testing the alternate hypothesis at a low significance level method, an important distinction for these calculations is that the p-values obtained are applied to the alternative hypothesis. This concept was briefly described in section 3.4.2 when describing the nature of this type of futility analysis method, but it is worth reiterating in order to fully understand the aggressive and conservative stopping rules. In general, for superiority studies investigating treatment effects between two or more therapies in a clinical trial setting, the p-value (significance level) calculations are

performed on the null hypothesis. They investigate the probability of whether the observed treatment difference can be attributed simply to chance, or to a real difference in the arms of study (i.e. treatment exposures). In other words, what is the probability that the null hypothesis is true, given the observed data at the time of analysis. If the probability is very small (for example a p-value of less than 0.05), then we can conclude that the difference is most likely due to treatment effect, and less likely to chance. These results favour the alternative hypothesis. Conversely, if the p-value is large enough, then we cannot effectively eliminate the effect of chance from our results, and therefore cannot conclude that there was a significant treatment effect. We must therefore accept the null hypothesis.

Testing for futility is very similar, but the difference is that the investigation is done on the alternate hypothesis instead of the null. It asks the question, what is the probability that the alternate hypothesis is 'true', given the observed data available. Again a p-value is obtained, but this information is used to decide if the trial should stop for lack of efficacy (reasons of futility), or if further investigation is warranted. If the p-value in this case is small (as in less than 0.005 with 25% of the data, or less than 0.02 with 50 – 75% of the data), then one could conclude that the trial should close, because there is very little evidence to support the alternative hypothesis. There is a significantly small probability of rejecting the alternative hypothesis when it is true, given the data provided. In this case, it is unlikely that the trial will demonstrate a positive outcome, and therefore a very small probability of successfully rejecting the null hypothesis. In a conservative setting, if the lower boundary of the observed HR >1 (i.e. treatment arm is worse than standard), then the same conclusion can be reached: that the trial could close early because of a very small probability of rejecting the null hypothesis. Conversely, if the p-value obtained at futility analysis is higher than the thresholds described, then there is not enough evidence to reject the alternative hypothesis, and therefore the trial should continue in order to obtain more information. A summary of the stopping rules are provided in the table below:

Table 5: Futility Analysis Stopping Rules

Time Point	Conservative Stopping (Low Significance)	Aggressive Stopping (Low Significance)	Conservative Stopping (Stochastic Curtailment)	Aggressive Stopping (Stochastic Curtailment)
25%	Stop if HR > 1	Stop if $p < 0.005 \mid H_a$	Stop if CP <0.1 (10%)	Stop if CP <0.2 (20%)
50%	Stop if HR > 1	Stop if $p < 0.02 \mid H_a$	Stop if CP <0.1 (10%)	Stop if CP <0.2 (20%)
75%	Stop if HR > 1	Stop if $p < 0.02 \mid H_a$	Stop if CP <0.1 (10%)	Stop if CP <0.2 (20%)
Note: HR = Hazard Ratio, Ha = Alternative Hypothesis, and CP = Conditional Probability				

3.4.5 Bootstrapping – Repeated Analysis at each Time point:

In order to increase the robustness of the investigation of futility analysis methods, and in order to limit the possibility of obtaining inaccurate or improbable futility analysis findings based on chance alone, bootstrapping methods for repeated sampling will be utilized. The intention is to gain insight into the variability of futility analysis results by providing a distribution of results, and to account for potential distortions caused by a specific sample that may not fully represent the overall trial sample (and therefore the population of interest). Bootstrapping refers to taking an existing sample or population of interest and re-sampling a number of times in order to obtain an alternate version of the single statistic of interest that would have been calculated from the original sample (Efron 1986, Efron 1993). Essentially, the process works by a program randomly selecting a sample from the total sample, and adding it to the pool that is being analyzed. This is done repeatedly until the required number of samples for analysis is achieved. The whole process can then be repeated to provide a new pool for analysis, as many times as needed. An important feature is that the sample selected is done completely at random, and the same sample can be selected more than once in each pool. After re-sampling a number of times, a distribution of the test statistic is formed, and from this mean, variance, confidence intervals, etc. can be obtained.

In this study, the original samples will be the OV.16 and CO.17 datasets. These two samples will be re-sampled repeatedly using bootstrapping methods, at the specific time points described above, in order to obtain a more robust description of futility analysis findings at each time point for each trial. Each trial will be re-sampled 1000 times at each retrospective futility analysis time point to ensure that chance findings are attenuated, as per previous studies using bootstrapping tactics at NCIC CTG (Ng 2007, Mittman 2009, Bradbury 2010). For each bootstrapping sample at each time point, the entire OV.16 or CO.17 study population will be available for selection.

3.4.6 Key Outcomes - Comparison of Results:

By bootstrapping the original trial samples and providing a report statistic of significance with each resample, a number of findings will be obtained at each specific time point, and for each futility analysis methodology. Test statistics for this study will include the futility analysis statistics (either conditional

probability for the stochastic curtailment method, or a p-value for the testing the alternative hypothesis at low significance level method), and these will be obtained from the original trial samples, and for each bootstrap sample of the original datasets. Final output statistics will also be provided for each original sample and for each bootstrap sample, and will include an overall p-value, hazard ratio, and confidence intervals. From this accumulation of results, histograms and scatterplots can be created to demonstrate the distributions of results, and averages of each statistic can be obtained for each interim analysis. A report of how many times the futility analysis would have met criteria for stopping will also be generated for each interim analysis (and in each setting), using a frequency table. These tables can then be analyzed using a chi square or Fisher exact test. These reports will be used to compare the results from each method at each interim analysis of futility.

In summary, futility analysis will be performed using two specific methods at 3 specific time points for each NCIC CTG trial, and using both aggressive and conservative stopping guidelines. In each case, bootstrapping will be used to increase the robustness of the investigation of the original trial datasets, and significance statistics will be obtained for each sample. The obtained results will be compared between futility analysis methods and with the final analysis findings of each trial, and reported accordingly. Distributions of averaged final analysis results from bootstrapping samples will also be compared to the results obtained from the original trial samples. The following tables provide a brief summary of the methods to be used:

Table 6: Summary of Study Methods

OV.16 Aggressive Setting				
Futility Analysis Method	Decision	25%	50%	75%
Low Significance	Continue			
	Stop			
Stochastic Curtailment	Continue			
	Stop			

CO.17 Aggressive Setting				
Futility Analysis Method	Decision	25%	50%	75%
Low Significance	Continue			
	Stop			
Stochastic Curtailment	Continue			
	Stop			

OV.16 Conservative Setting				
Futility Analysis Method	Decision	25%	50%	75%
Low Significance	Continue			
	Stop			
Stochastic Curtailment	Continue			
	Stop			

CO.17 Conservative Setting				
Futility Analysis Method	Decision	25 %	50%	75%
Low Significance	Continue			
	Stop			
Stochastic Curtailment	Continue			
	Stop			

3.5 Ethical Considerations:

There are very few ethical considerations for this investigation. The original data from the trials being used for this study have been previously published, and no further information is required from the patients. The intent of this study is to examine futility analysis methodologies, which has very little consequence on previously submitted patient data. However, patient data that has been previously collected for the purposes of OV.16 and CO.17 objectives will be used for this investigation, and therefore this study required ethics approval. The secondary analysis proposed for this study will not include any information that can be used to identify the patients who participated in the associated trials. This study was submitted to the Queens University Ethics Review board for expedited approval on January 20, 2012, and approval was obtained on the 6th of February, 2012. Approval was also obtained at NCIC CTG to use the data and the previously published summary tables for both trials.

3.6 Results:

3.6.1 Data Analysis:

Once approval for use of the data was granted, a SAS file was provided for both OV.16 and CO.17 trials. The SAS files contained a copy of the final and complete data sets for each trial. The data set contained the raw data acquired during the course of the trials, which had been previously reviewed, cleaned, and collated for the original final analyses of each trial. The first step for the purposes of this project was to run the final analysis on each data set, in order to ensure that identical results could be achieved. This step would serve to validate, not only the data provided, but also the program developed to obtain the final statistics: hazard ratio, p value, confidence intervals, final sample size, and number of events. The median survival and disease free survival times were also obtained, and compared to the previously released abstracts, in order to ensure consistency. In order to perform this validation, the parameters for each trial were inputted into the program (e.g. stratification factors, target statistics, expected events and sample size, etc.), and the final analysis was repeated using the data obtained during the trial. Once confident that the data sets were complete and consistent, and the SAS programs

for obtaining the desired output statistics were accurate, the next step was to determine the actual time points to run the retrospective interim analyses on each trial.

In order to specify a date for each interim analysis, the data sets were ordered chronologically by primary endpoint dates: date of progression for OV.16 and date of death for CO.17. Next, quartiles for the total required number of events for final analysis were calculated in order to determine the actual number of events required at 25%, 50%, and 75% time points: 158, 316, and 473 respectively for OV.16, and 111, 223, and 334 for CO.17. With the data set organized chronologically, and the 'cut off' numbers obtained (i.e. number of events needed to initiate each retrospective interim analysis), the associated dates for the relevant n^{th} event could be easily identified, and used for the specific cut off time point of each retrospective interim analysis. For OV.16, the 158^{th}, 316^{th}, and 473^{rd} events occurred on 2004-FEB-24, 2005-JAN-12, and 2005-OCT-19. For CO.17, the dates used were 2004-NOV-30, 2005-MAY-05, and 2005-SEP-05, which corresponded to the dates of the 111^{th}, 223^{rd}, and 334^{th} events. These dates were used in the SAS programs for the retrospective futility analyses. The time of final analysis in this project (i.e. 100% time point) for both OV.16 and CO.17 was 31-MAR-2008, when the trials had both met their total required events needed for their primary final analysis.

Next, SAS programs were written for obtaining futility analysis statistics at each time point, and for each trial. A SAS macro was used to obtain a significance statistic (p-value for the alternative hypothesis) using the 'testing the alternative hypothesis at a low significance' method, as well as a conditional probability statistic using stochastic curtailment. This was done using the Jennison and Turnbull formula (Jennison 2000). Again, the same initial parameters for each trial were inputted into the SAS program, along with the retrospective futility analysis cut off dates for each time point. The program was run, and the desired 'futility analysis decision' statistics, in addition to the previous output statistics (HR, number of events, sample size, p-value, and confidence intervals), were obtained. The program was run for each trial, which produced one single output for each trial containing the desired statistics at each time point, and this was reviewed to ensure the program functioning appropriately before carrying on to adding repeated samples (bootstrapping) parameters.

Once confident with our SAS program and its output, the next step was to increase the number of results, or samples, for each interim futility analysis. In order to accomplish this, a macro was created containing the parameters required for bootstrapping. The total number of repeated samples was specified as 1000 for each

retrospective interim futility analysis, and a do loop was created for the output statistics. This updated program, now containing initial study parameters, interim analysis cut off dates, output statistics including futility analysis results and overall statistics of trial significance, and now bootstrapping/repeated sample parameters, was applied to the initial data sets. The result was a massive data set containing 1000 sample results for each interim analysis (4000 per trial, 8000 total). This entire data set was then exported to an excel spreadsheet for further manipulation.

Once the data was exported to excel, a table was generated. The table headers were labeled, and filters were applied in order to view data by interim analysis time point. Next, new columns were created in the table to summarize the decision of each repeated sample in terms of the futility analysis finding using either method. That is to say, a new column was created, and a new formula was introduced, which would summarize the individual results (produced by repeated analysis/bootstrapping techniques) for each method of futility analysis: to the p-value statistic obtained from the 'testing the alternative hypothesis at a low significance level' method, and to the conditional probability statistic for the stochastic curtailment method. The formulas used to summarize the 'go/no go' results were derived using both aggressive and conservative stopping rules. These summary columns could then be filled with the futility analysis results, and tabulated in order to produce a frequency table, which could compare the two futility analysis methods. Each table would have 1000 total samples, and therefore 1000 results for deciding whether a trial should 'continue' or 'stop'. Again, the filters would provide the ability to view the total number of decisions (stop or continue) of each futility analysis method in each setting or time point (25%, 50%, 75% or 100%). The results of these frequency tables could then be analyzed using chi square and Fisher exact tests.

Averages for HR, p-value, Confidence Intervals, and number of events were also calculated from the bootstrap samples, which could then be compared to the actual results obtained from each clinical trial's primary analysis. The process described above was used for both clinical trials. An example of the final output can be seen in the table below:

Table 7: OV.16 SAS Output – Futility Analysis

OV.16 SAS OUTPUT - FUTILITY ANALYSIS													
HR	pvalue	bootnu	TIMING	CI1	CI2	events	cPower	Decision (CP)-Conservati	Decision (CP)-Aggressive	pRejH1	Decision (p-value) - Cons	Decision (p-value) - Aggressive (25%)	Decision (p-value) - Aggressive (50%, 75
1.50249	0.02554	1	25%	1.05107	2.1478	120.351	0.22915	Continue	Continue	0.00702	Stop	Continue	Stop
1.27153	0.05512	1	50%	0.99475	1.62531	254.988	0.02529	Stop	Stop	0.01613	Continue	Continue	Stop
1.35587	0.00334	1	75%	1.10641	1.66159	371.671	0	Stop	Stop	0.00079	Stop	Stop	Stop
1.25441	0.00606	1	100%	1.06696	1.4748	586.537	0	Stop	Stop	0.0015	Stop	Stop	Stop
1.44598	0.09274	2	25%	0.94067	2.22275	83.12	0.44169	Continue	Continue	0.02844	Continue	Continue	Continue
1.20386	0.16917	2	50%	0.92408	1.56835	219.655	0.12113	Continue	Stop	0.05502	Continue	Continue	Continue
1.19314	0.10034	2	75%	0.96653	1.47289	346.344	0.0022	Stop	Stop	0.031	Continue	Continue	Continue
1.07183	0.3973	2	100%	0.91279	1.25857	595.669	0	Stop	Stop	0.1424	Continue	Continue	Continue
0.84884	0.4165	3	25%	0.57166	1.26042	98.317	0.7855	Continue	Continue	0.72219	Continue	Continue	Continue
1.17182	0.24531	3	50%	0.89679	1.53119	214.756	0.16567	Continue	Stop	0.08303	Continue	Continue	Continue
1.23465	0.04749	3	75%	1.00234	1.5208	353.632	0.00043	Stop	Stop	0.01372	Stop	Continue	Stop
0.94742	0.51743	3	100%	0.80449	1.11575	574.557	0.00013	Stop	Stop	0.66428	Continue	Continue	Continue
0.92143	0.66652	4	25%	0.6351	1.33687	110.944	0.72054	Continue	Continue	0.5823	Continue	Continue	Continue
0.87306	0.31541	4	50%	0.66981	1.138	218.786	0.71622	Continue	Continue	0.78254	Continue	Continue	Continue
1.02556	0.80593	4	75%	0.83851	1.25435	378.924	0.05146	Stop	Stop	0.3196	Continue	Continue	Continue
1.17463	0.05178	4	100%	0.99876	1.38148	584.068	0	Stop	Stop	0.01508	Continue	Continue	Stop

3.6.2 Results from Original Data Sets

The following table describes the results obtained when applying the retrospective futility analysis dates to the original data sets, and executing the programming for futility analysis and final analysis statistics:

Table 8: Original Data Results – OV.16 and CO.17

OV.16 SAS OUTPUT - FUTILITY ANALYSIS							
TIMING	HR	p-value	CI1	CI2	events	cPower	pRejH1
25%	1.1	0.58	0.78	1.57	123	0.52	0.22
50%	1.17	0.22	0.91	1.49	257	0.08	0.07
75%	1.21	0.06	0.99	1.47	408	0	0.017
100%	1.1	0.25	0.94	1.28	621	0	0.08
CO.17 SAS OUTPUT - FUTILITY ANALYSIS							
TIMING	HR	p-value	CI1	CI2	events	cPower	pRejH1
25%	0.94	0.75	0.63	1.39	98	0.44	0.52
50%	0.74	0.028	0.56	0.97	207	1.42	0.97
75%	0.63	0.00007	0.5	0.79	304	3.87	1
100%	0.77	0.005	0.64	0.92	450	7.58	0.99

As can be seen from the table above, retrospective futility analyses and final analysis investigations were performed at the time that 25%, 50%, and 75% of the total required events were obtained for each trial. For OV.16, the hazard ratio stayed above 1, and remained similar to the eventual final analysis result of 1.1. The confidence intervals for the hazard ratio contained 1 as well throughout the investigation. The p-value (applied to the null hypothesis) for this trial remained above the 0.05 mark at each interim analysis as well, however at the 75% analysis the p-value was smallest at 0.06. A closer investigation of the data, however, suggests that this 'nearly significant' p-value is actually demonstrating a potential significant difference in favour of the control arm. This means that at the 75% analysis, the data were actually favouring the control arm for efficacy, and the experimental/investigational arm, therefore, was actually performing worse than the control. Once all data was obtained, the data shows that the trial was unable to achieve its objectives, could not show a significant difference in treatment effects of the two arms of study, and could not reject the null hypothesis (HR = 1.1, p-value = 0.25). These findings were as expected at the onset of this current study. What is of particular interest in this investigation, however, are the differences in the results of the futility analysis methods for OV.16.

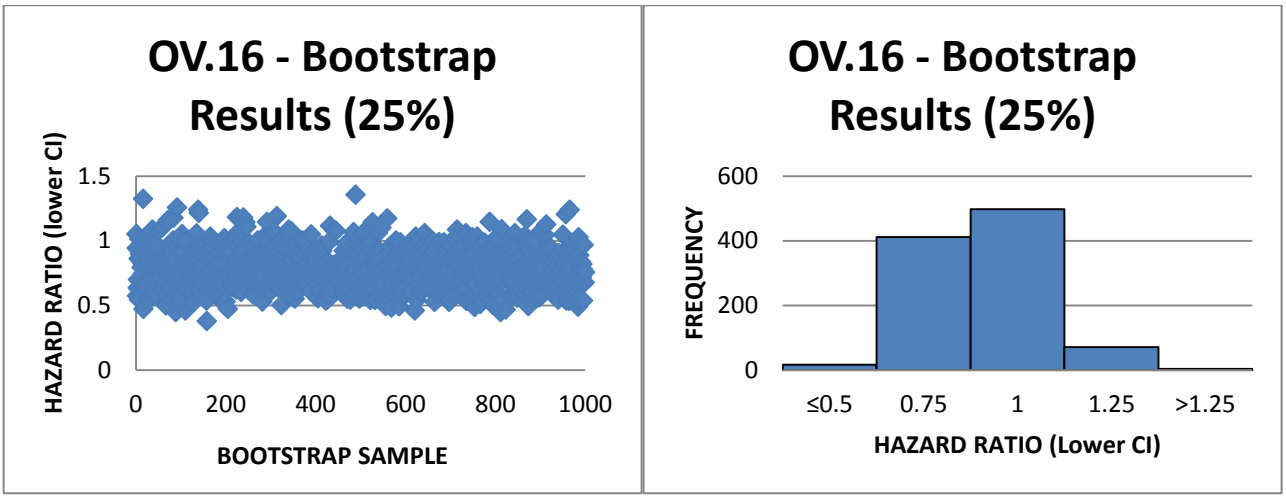
Conditional power values using the stochastic curtailment method are labeled as 'cPower' in Table 8. These values show that the trial could have been stopped early for lack of efficacy (futility) as early as the 50% analysis, where the conditional power value was 0.08. This means that with 50% of the data acquired, there is only an 8% probability of obtaining a significant result at final analysis. With 75% of the data acquired, this value is very close to 0. Recall that at this analysis, the treatment differences were in favour of the control arm. At both of these interim analyses, a decision to stop the trial could have been reached using either aggressive or conservative stopping rules. The results for the testing the alternative hypothesis at a low significance level, however, differed from the stochastic curtailment method. The p-value for this method (applied to the alternative hypothesis) is labeled as pRejH1 in Table 8. The only value obtained from this investigation that could warrant a decision to close the trial early can be seen at the 75% analysis (0.017), but this would only provide evidence to close the trial early using aggressive stopping rules. The futility analysis could not provide sufficient evidence to stop the trial early at 75% using conservative stopping, as the lower confidence intervals for hazard ratio was not above 1. At this point, however, it was very close (CI = 0.99, 1.47), and perhaps a DSMC may have considered early termination.

CO.17, as expected, produced very different results from OV.16. The hazard ratio for this trial remained below 1, as did the confidence intervals for all investigations except the 25% analysis. After the 25% analysis, the overall p-value was significant, and the largest treatment difference was seen at the 75% analysis. At this analysis, the hazard ratio was 0.63, the p-value was 0.00007, and the largest difference in treatment arms can be seen in Figure 2. The final analysis demonstrates the treatment effect in favour of the investigational arm (cetuximab arm), with a hazard ratio of 0.77, and p-value of 0.005, and confidence intervals below 1. The futility analysis methods, in this case, were consistent. Both methods could not provide evidence to stop the trial early for lack of efficacy, and this too was as anticipated at the onset of this investigation. The differences in futility analysis results seen in OV.16, however, were not anticipated at the onset of the study. This merits further investigation, and therefore bootstrapping methods will be applied to increase the number of samples in order to explore these findings further.

3.6.3 Bootstrap Samples – Description of Results:

The following figures describe the distribution of results for the lower boundary of HR by bootstrapping for OV.16, a trial which in the final analysis did not demonstrate significant findings.

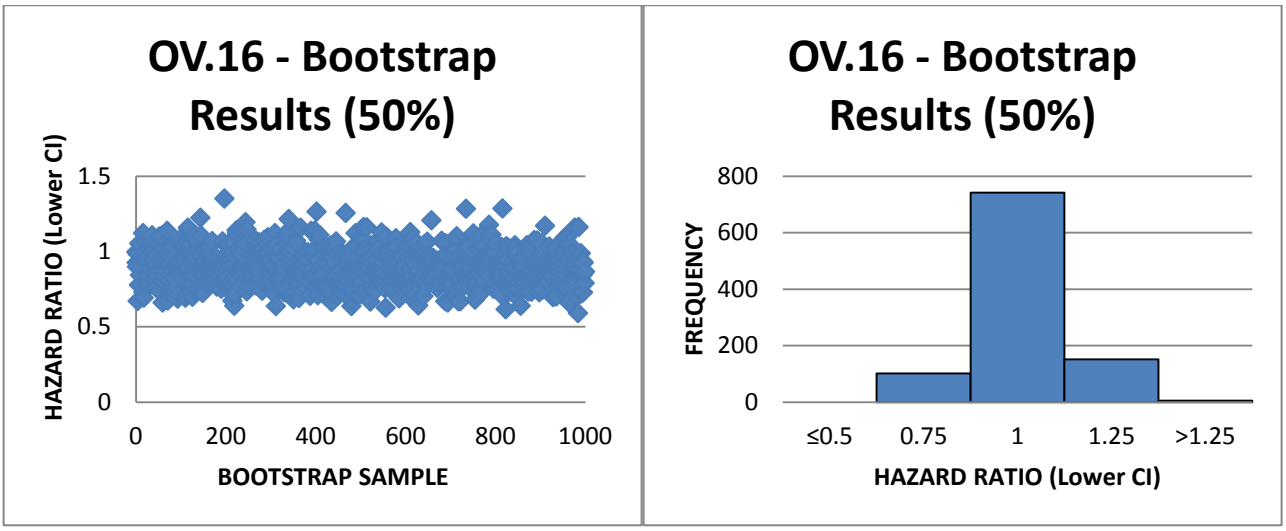
Figure 3: OV.16 Bootstrap Results (25%)



With approximately 25% of the required number of events acquired, 106/630 events on average, an analysis of the hazard ratio shows considerable variability. The confidence intervals for this interim analysis are very large (average CI = 0.79, 1.68), and the lower boundary is much more variable. It is extremely difficult to

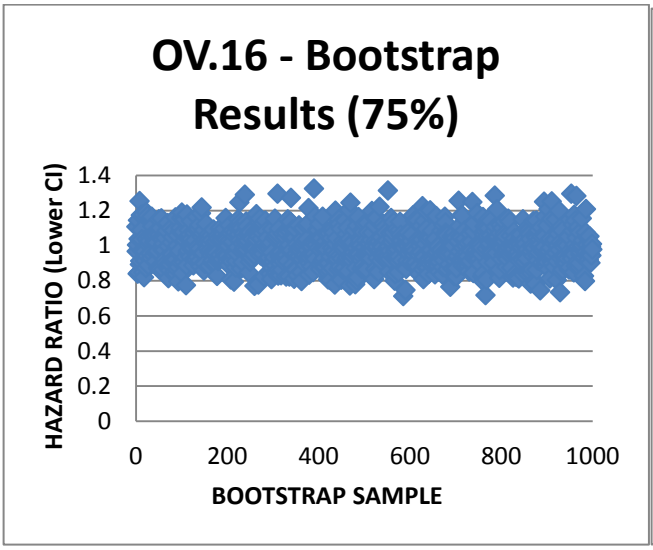
determine if this trial would reach significance from the data acquired at this point, as can be seen in the futility analysis results of 3.6.4. Overall, however, the average HR at this time point is 1.15 (median 1.13), which could suggest that the trial might not provide final results significantly favouring the alternate hypothesis. Furthermore, the average p-value for this interim analysis is 0.44 (median 0.4), which provides further evidence that we would not be able to reject the null hypothesis at final analysis. The statistics for the original data set at this point were similar to the average from bootstrap samples (95% CI = 0.78, 1.57, HR = 1.1, p-value = 0.58).

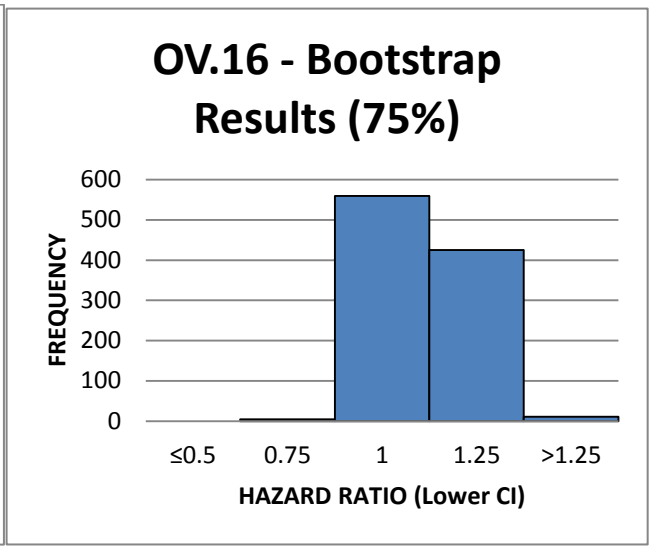
Figure 4: OV.16 Bootstrap Results (50%)



With 50% of the required data acquired, approximately 227 events, an interesting trend is developing. The variability in the hazard ratio is becoming smaller, and therefore the confidence intervals are becoming more narrow (average CI = 0.89, 1.5). Here, the lower boundary of HR is starting to approach 1, although the overall average HR has changed very little at 1.16 (median 1.15). The average p-value is now 0.36 (median 0.28), but the result is still far from the desired threshold needed to determine a 'significant effect' (i.e. to reject the null hypothesis). Again, the original data set values are similar to the averages of the bootstrap samples (95% CI = 0.91, 1.49, HR = 1.17, p-value = 0.21).

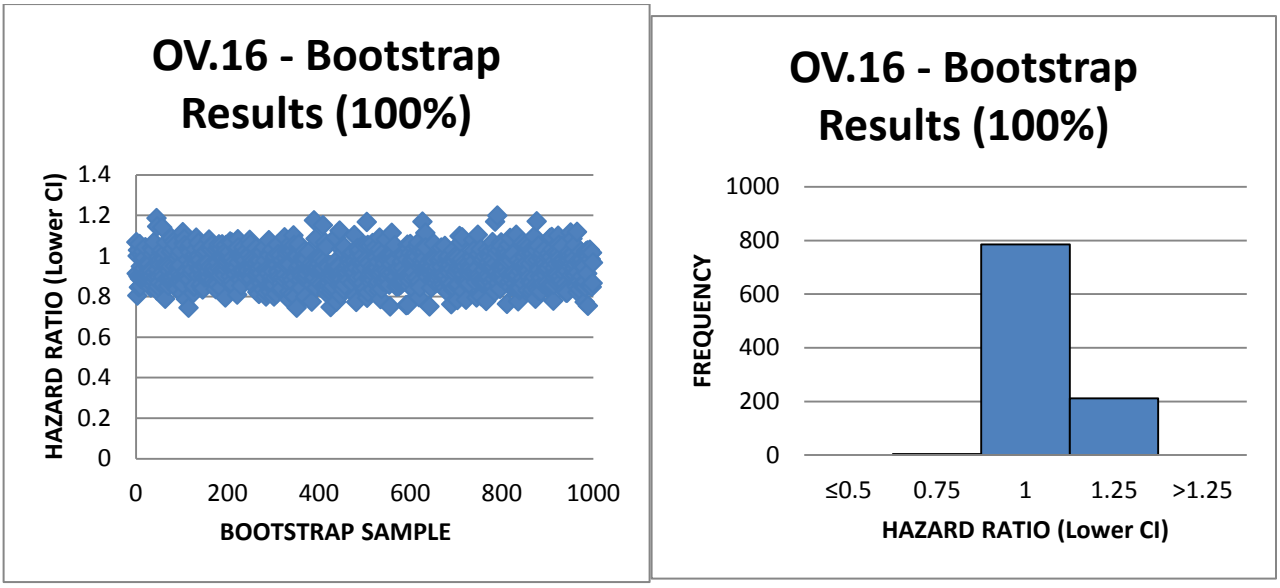
Figure 5: OV.16 Bootstrap Results (75%)





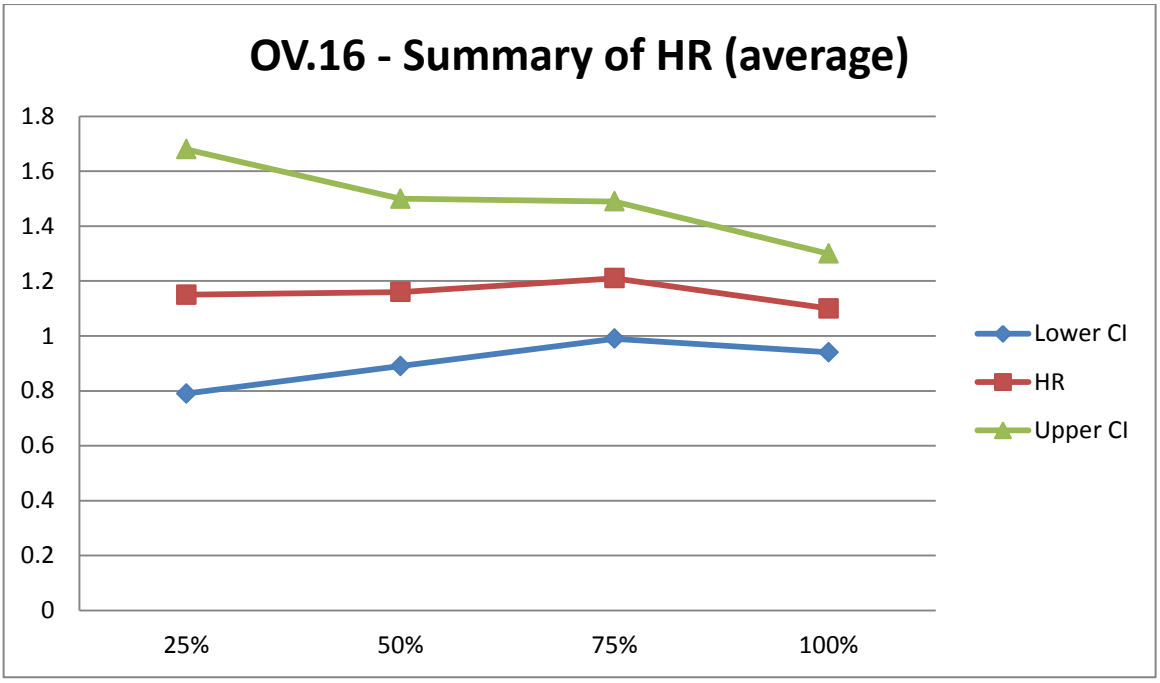
At 75%, or approximately 369 events, one can start to extrapolate how this trial will most likely conclude. The trend in the confidence intervals continues, as the average lower boundary continues to approach 1, and the variability between high and low boundaries becomes even smaller (average CI = 0.99, 1.49). Overall, the average HR is now 1.21 (median 1.2), and the p-value has become smaller at 0.18 (median 0.07). This information, at this particular time point, is suggesting that the experimental arm may in fact be inferior to the control (as was seen in the original data analysis). While these results are not significant, this information would have been troubling for a DSMC reviewing the interim data at this point, as the probability of demonstrating a positive trial given the information to date would have been low. Again, the original data set values are similar to the averages of the bootstrap samples at this analysis (95% CI = 0.99, 1.47, HR = 1.21, p-value = 0.06).

Figure 6: OV.16 Bootstrap Results (100%)



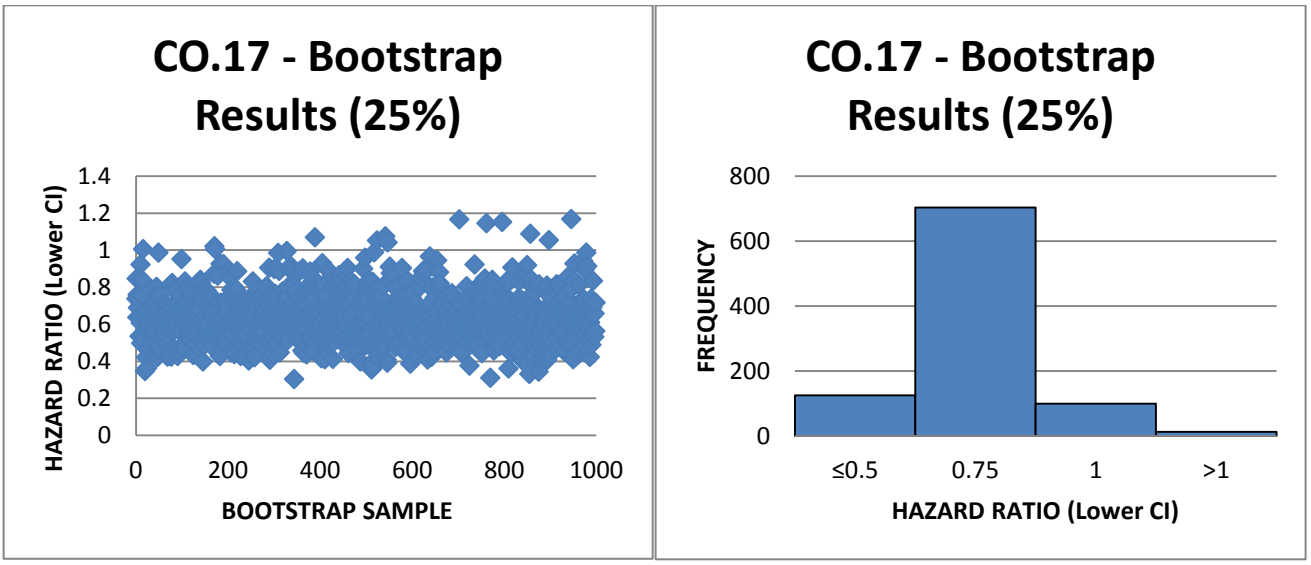
At the time of the final analysis, when the total required number of events had occurred, the confidence intervals on average were 0.94 and 1.3. The overall average hazard ratio was 1.1, and the p-value was 0.33. From this data, we can conclude that the trial could not provide sufficient evidence to reject the null hypothesis. The trends seen throughout the retrospective interim analyses appear to have correctly predicted the outcome of the trial, in this particular case. Further, the averaged results provided by bootstrapping were very similar to the actual final analysis results of this trial: recall HR = 1.10, 95% CI = 0.95 – 1.30, p-value = 0.25. Figure 7 provides a summary of the hazard ratio findings, and relative confidence intervals, at each retrospective interim analysis acquired from the bootstrapping samples. The average HR values from the boostrap samples are mapped at each time point in the table below:

Figure 7: OV.16 Summary of HR



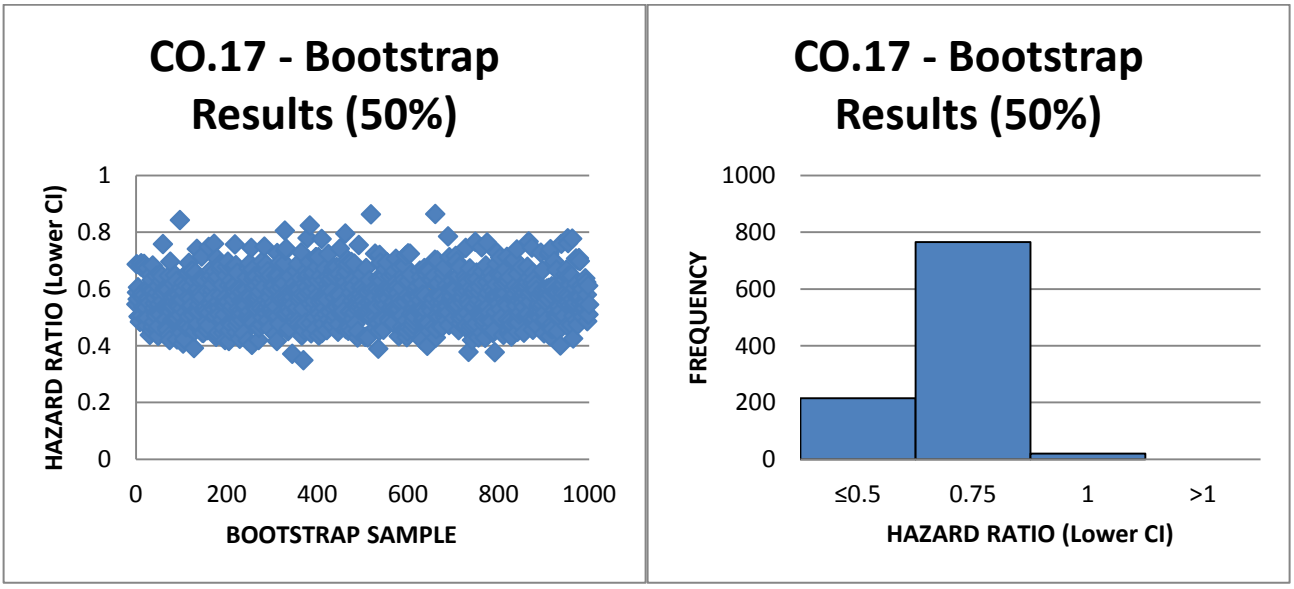
In contrast to OV.16, CO.17 did in fact demonstrate significant evidence to reject the null hypothesis. The following figures describe the distribution of the results from bootstrapping of the lower boundary of HR for this trial.

Figure 8: CO.17 Bootstrap Results (25%)



With approximately 25% of the required events acquired, or 93 of 445 deaths on average, similar trends appear from bootstrapping as in OV.16. The confidence intervals are very wide (average CI = 0.64, 1.44), and there is considerable variability in the lower boundary of HR. The average lower CI, however, is much lower for this trial then was the case in OV.16. The overall average HR at this point is 0.96 (median 0.94), and the p-value is 0.5 (median 0.49), which shows that there is not much evidence to support or reject the alternate hypothesis with the data available at the time of this interim analysis. The original data set values for CO.17 are similar to the averages of the bootstrap samples (95% CI = 0.63, 1.39, HR = 0.94, p-value = 0.75).

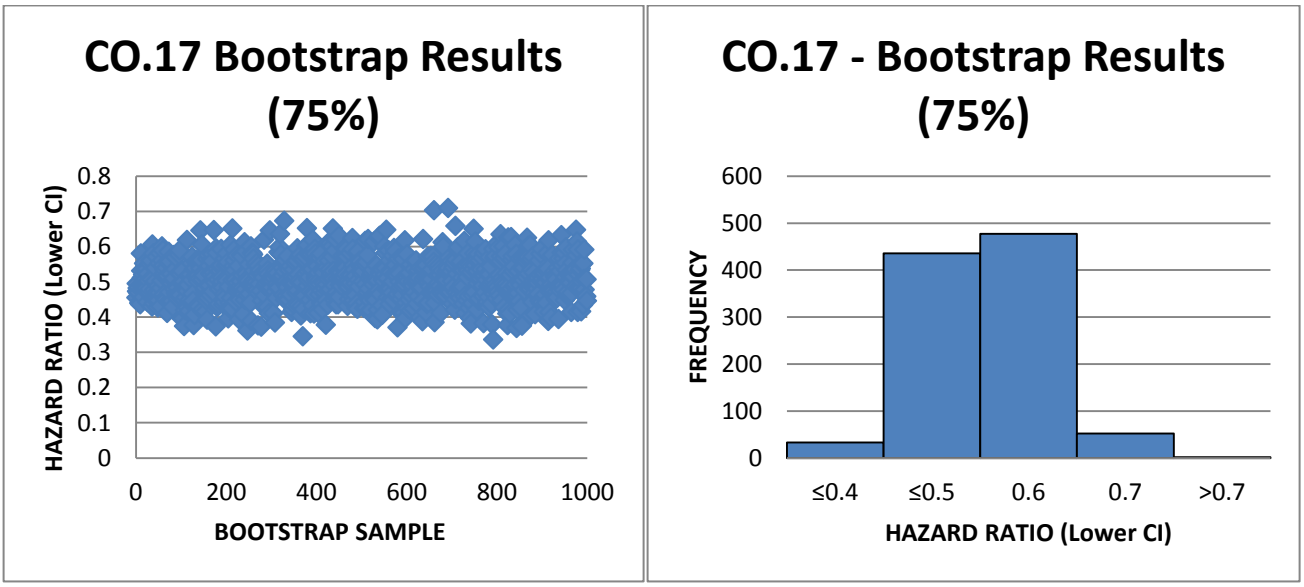
Figure 9: CO.17 Bootstrap Results (50%)



At the 50% time point, when on average 201 events had occurred, a similar trend is occurring as with OV.16, but the result is pointing to a different outcome. Although there still exists considerable variability in the confidence intervals (average 0.57, 0.99), the lower boundary of the HR appears to be moving away from 1, as is the overall average HR. In addition, as can be seen from the graph above, all values of the lower boundary of HR are below 1. The average overall HR is now 0.75, and the p-value is approaching significance at 0.13. It would appear from the data available, that a trend towards significance is establishing, and that this trial could potentially provide sufficient evidence to support the alternative hypothesis at the final analysis. Again, the original data set values for CO.17 are similar to the averages of the bootstrap samples (95% CI = 0.56, 0.96, HR

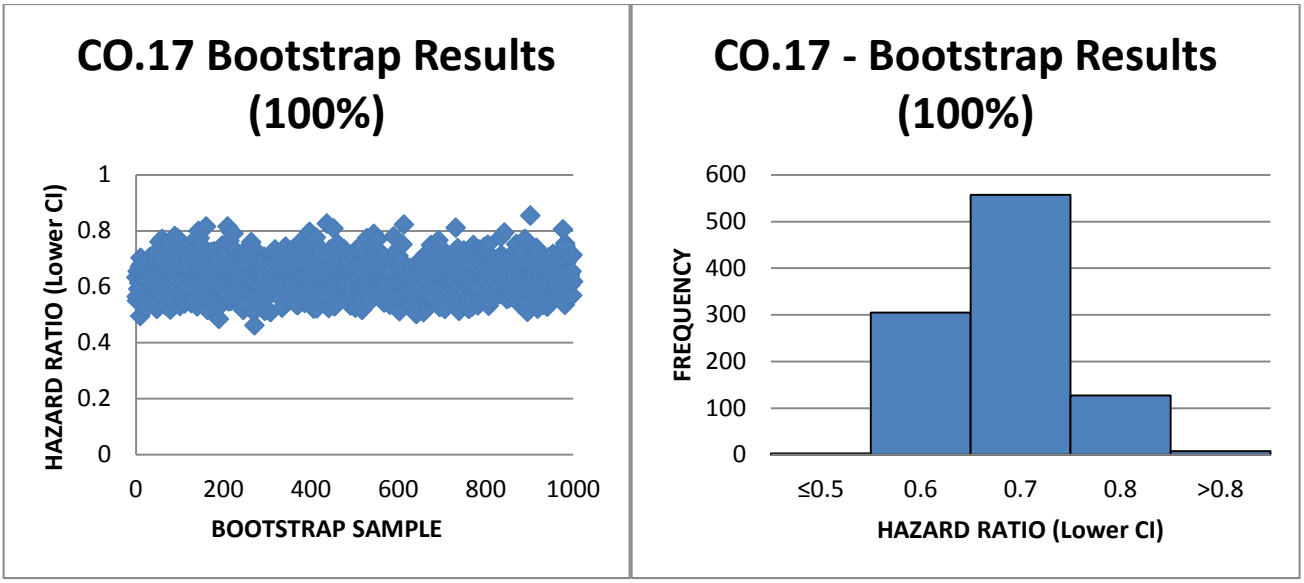
= 0.74) at this point, with the exception of the p-value, which in the original data set is 0.028 and is 0.13 in the averaged results from the bootstrap samples.

Figure 10: CO.17 Bootstrap Results (75%)



With 75% of the data acquired (average 296 events), the trend continues as there is now convincing evidence to support the alternative hypothesis. At this point, the confidence intervals are quite narrow (CI = 0.5, 0.8 on average), and the lower boundary of the HR is quite low at 0.5. Again, none of the values for the lower boundary are even above 0.8 at this point. Furthermore, the average hazard ratio is now 0.63, and the p-value is 0.003. Had this interim analysis included an analysis of benefit, one could speculate that the DSMC might have considered closing for benefit at this point (efficacy in favour of the experimental arm). The original data set values for CO.17 are once again similar to the averages of the bootstrap samples (95% CI = 0.5, 0.8, HR = 0.63) at this point, and again with the exception of the p-value, which in the original data set is 0.00007 and is 0.003 in the averaged results from the bootstrap samples. Though, they are significant in each case.

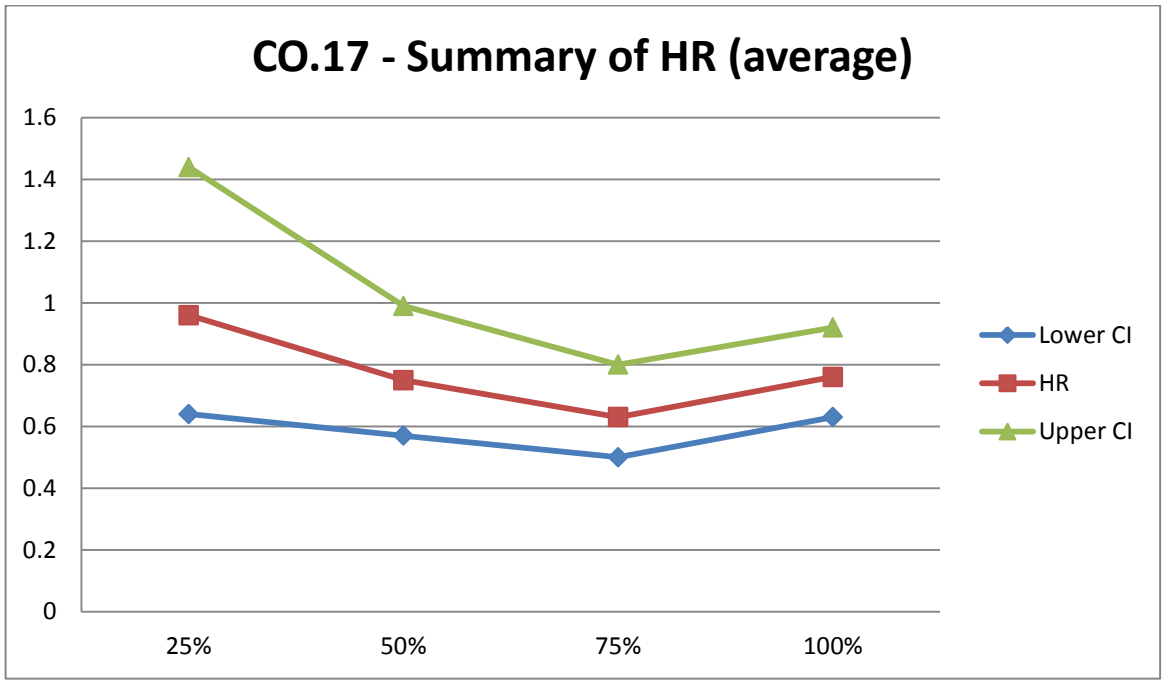
Figure 11: CO.17 Bootstrap Results (100%)



Once all of the data had been accumulated, a peculiar finding has occurred. The trial, while still considered positive, as it showed significant results and could therefore conclusively reject the null hypothesis, has actually concluded with results that were less significant than those seen at the 75% analysis. At the 75% interim analysis, the hazard ratio shows more of a benefit for the experimental arm than at the final analysis (0.63 versus 0.76 here). The p-value, while still significant at the final analysis, was also smaller at the 75% analysis (0.003 versus 0.04). The trend that was seen from the three interim analyses of futility did prove to correctly predict that the trial would finish with significant results in favour of the experimental arm, but the trend did not continue into the final analysis. The figure below (Figure 12) summarizing the interim result findings for CO.17 shows this phenomenon in the overall survival of the trial sample. This curve is sometimes described as what is called a 'banana curve'. While the average overall hazard ratio in OV.16 remained relatively similar at each analysis, in this trial there is considerable variability in the average overall HR at each interim analysis. This is interesting because, had the trial been stopped earlier, the results would have been more suggestive of benefit for the experimental arm. The results appear to even out towards the end of the trial as more events start to occur on both arms of treatment. At the final analysis, the confidence intervals have changed from the previous analysis (CI = 0.63, 0.92), but, like the interim analysis at 75%, they do remain quite narrow. As expected, the trial was considered to have a positive result, and the findings were published. The result was an increase in median survival time from 4.6 to 6.1 months. Again, the averaged results provided by

bootstrapping were very similar to the actual final analysis results of this trial: recall HR = 0.77, 95% CI = 0.64 – 0.92, P = 0. 005. Below is a summary of the hazard ratio findings, and relative confidence intervals, at each retrospective interim analysis:

Figure 12: CO.17 Summary of HR



3.6.4 Futility Analysis:

Much like the trends observed in the analysis of the endpoint parameters described in section 3.6.3, as well as the analysis the analysis of original data in section 3.6.2, some interesting findings were also observed when investigating the two different futility analysis methods in both aggressive and conservative settings, using the bootstrap samples. The two tables below summarize the results of the futility analysis methods used for retrospective interim analysis for the CO.17 trial. Recall that there were 1000 samples at each retrospective analysis by bootstrapping, and that each analysis provided an outcome parameter, or test statistic, for each futility analysis method. These outcome parameters provide evidence to support a decision as to whether the trial should be allowed to continue or stop based on findings of efficacy (or lack of efficacy), based on the data acquired at the time of each analysis. The decisions to continue or stop a trial can be used by applying either aggressive or conservative rules for truncating a trial (hence the terminology 'conservative or aggressive settings'). Table 9 provides a summary of these findings for the CO.17 trial:

Table 9: CO.17 Futility Analysis Results

CO.17 - Conservative Setting					
FA Method		Decision	25%	50%	75%
A	Testing H_a at Low α (Stop if HR >1)	CONTINUE	987	1000	1000
		STOP	13	0	0
B	Stochastic Curtailment (Stop if CP <0.1)	CONTINUE	1000	998	1000
		STOP	0	2	0
CO.17 - Aggressive Setting					
FA Method		Decision	25%	50%	75%
A	Testing H_a at Low α (Stop if Ha < 0.005)	CONTINUE	994	1000	1000
		STOP	6	0	0
B	Stochastic Curtailment (Stop if CP <0.2)	CONTINUE	994	995	1000
		STOP	6	5	0

For CO.17, a trial that was consistently showing a difference in treatment arms in favour of the experimental arm which ultimately reached significance and resulted in a positive, there are very little differences in the results of the retrospective futility analysis findings. Both methods suggested that the trial should continue in an overwhelming majority of cases, and at each interim futility analysis. At the 25% analysis, a peculiarity is seen between the aggressive and conservative results for Method A (i.e. Testing the alternative hypothesis at a very low significance). Here, the rules would have seen the trial stop in 13 cases using 'conservative' rules, but in only 6 cases using 'aggressive' rules. This occurred in the samples that obtained the highest hazard ratios (average HR of these 13 samples = 1.63, whereas overall average HR at from all samples at 25% is 0.96), where the confidence intervals did not contain 1. The confidence intervals for these samples, however, were still very wide (average CI = 1.08, 2.46). Because of this large variability in the confidence intervals, in addition to the small amount of data acquired at this time, the aggressive rules did not flag as frequently since the aggressive rules were based specifically on the futility analysis statistics instead of the hazard ratio. The futility analysis statistics would take into account this variability due to lack of acquired evidence. This anomaly resulted in a significant difference between Method A and Method B in the conservative setting, but not in the aggressive setting for this 25% analysis (p-value = 0.0002 using Fisher exact test).

Furthermore, in the aggressive setting, we can see that both methods flagged 6 samples to stop. The methods, however, were only consistent in 4 of 6 of these samples (67%) This tells us that futility analyses at 25% are highly variable, and the results differ because of the volatility of the data at this early stage of a trial. In

contrast to at 25%, where Method A flagged the samples to close more often than Method B, we can see that the stochastic curtailment method flagged more samples to stop in both conservative and aggressive settings at the 50% analysis. In this case, the aggressive setting resulted in more decisions to stop, as would be expected. These results, however, were not significantly different from Method A results (p-value by Fisher exact test = 0.500 for comparison of methods in conservative setting, and p-value = 0.06 in aggressive setting). Of course, as can be seen from the results above, all samples suggested the trial should continue using either method at 75%.

We can see from these results that, with respect to CO.17 data, both futility analysis methods were accurate in determining that the trial should continue in the majority of cases, although variability between methods could be seen, and that the accuracy of these predictions increased as more data became available. However, there was a significant difference at the 25% analysis between futility analysis methods in the conservative setting, and some inconsistency in the results using conservative and aggressive stopping rules for the same method. These findings can most likely be attributed to the extreme variability of the data due to lack of sufficient evidence at this early analysis. Overall, analyses using both methods resulted in decisions to continue the majority of the time, and this was consistent with the analyses performed on the original CO.17 dataset, as outlined in section 3.6.2. More substantial differences, however, can be seen in the retrospective futility analyses for OV.16, as can be seen from the summary table below, Table 10.

Table 10: OV.16 Futility Analysis Results

OV.16 - Conservative Setting					
FA Method		Decision	25%	50%	75%
A	Testing H_a at Low α (Stop if HR >1)	CONTINUE	926	844	564
		STOP	74	156	436
B	Stochastic Curtailment (Stop if CP <0.1)	CONTINUE	1000	652	35
		STOP	0	348	965
OV.16 - Aggressive Setting					
FA Method		Decision	25%	50%	75%
A	Testing H_a at Low α (Stop if Ha < 0.005)	CONTINUE	972	804	507
		STOP	28	196	493
B	Stochastic Curtailment (Stop if CP <0.2)	CONTINUE	992	419	14
		STOP	8	581	986

As can be seen from the table above, similar trends are also observed from the futility analysis results of the OV.16 trial. In fact, these trends are better illustrated in this trial, which did not reach statistical significance in its final analysis. The results of futility analysis Method A (testing the alternative hypothesis at a low significance level) at the 25% analysis again shows that the conservative stopping rules flag the trial to close more often than in the aggressive setting (74 times versus 28), and again this difference is significant (p-value = 2.21×10^{-6}). This is largely due to the variability in the hazard ratios obtained from the bootstrap samples with so little data accumulated at this point, and also because the conservative rules does not account for this variability. Again, at this point, the stochastic curtailment method (Method B) does not flag the trial to stop as frequently as Method A in either conservative or aggressive settings. At this 25% analysis, however, both methods predominantly suggest the continuation the trial in the majority of samples. Under the conservative setting, there is again a significant difference between Methods A and B at the 25% analysis (p-value = 2.6×10^{-23}), but now this significant difference can also be seen in the aggressive setting as well (p-value = 0.0008).

Also unlike CO.17, in this trial, there is a trend towards suggesting trial closure from both futility analysis methods as more information becomes available. This is to be expected from a trial that could not show significant differences at final analysis, and therefore could not reject the null hypothesis. At the 50% analysis, however, the stochastic curtailment method had more cases where the trial reached the threshold for demonstrating futility, and therefore providing evidence to close the trial early for lack of efficacy, than that observed using Method A. This was observed in both conservative and aggressive settings as well. Method B produced 348 results to stop the trial in a conservative setting versus 156 in Method A (35% of cases versus 16% of cases, p-value = 2.0^{-23} by Fisher exact testing), and 581 versus 196 in an aggressive setting (58% versus 20%, p-value = 7.3^{-72}). This difference was even more apparent at the 75% analysis, and again in both settings, with a decision to stop the trial 965 and 986 times for Method B, versus only 436 and 493 times for Method A (97% and 99% using Method B versus 43% and 49% using Method A). The p-values at the 75% analysis were 2.4×10^{-168} in the conservative setting when comparing Method A and B results, and 1.6×10^{-166} in the aggressive setting.

These findings suggest that, while both methods tend to become more consistent with final analysis results as more data accumulates, Method B appears to be a much better predictor of trial futility. This method

was most accurate and consistent with 50% and 75% of the data acquired, and conclusively would have shown that the trial should not continue for reasons of futility, based on the data available. Again there was significant variability at the 25% analysis, which made comparisons difficult to interpret, and would suggest that it would be difficult to accurately assess whether a trial should continue or not with this little data available. These findings are consistent with the futility analyses performed on the original OV.16 dataset, where the stochastic curtailment method provided evidence to stop the trial early at the 50% and 75% analyses using either aggressive or conservative stopping rules. Testing the alternative hypothesis at a low significance, however, provided evidence to close the trial at only the 75% analysis, and only using aggressive stopping rules. Again, these findings were not anticipated at the onset of this investigation.

3.6.5 Conservative versus Aggressive Settings:

As alluded to in the previous section, there were also differences observed when comparing the same futility analysis method in both aggressive and conservative settings. Because of the variability in results at 25%, and the volatility of the data this early in the trial, these differences at this time point will not be discussed further in this section. However, differences could still be observed when comparing the stopping rules used for futility analysis at the 50% and 75% time points, but this could only be seen in the OV.16 results. In this trial, the results were significantly different when comparing conservative stopping results with aggressive stopping results for both futility analysis methods. Method A showed more decisions to close the trial early using aggressive stopping rules at the 50% (196 versus 156, p-value = 0.006) and 75% analyses (493 versus 436, p-value = 0.002), than was seen with the conservative stopping rules. Method B also showed more decisions to close the trial using aggressive rules versus conservative rules as well in both the 50% analysis (581 versus 348, p-value = 9.1^{-26}), and at the 75% analysis (986 versus 965, p-value = 0.002). These findings were not expected at the onset of this investigation.

CHAPTER 4 –Conclusions and Discussion:

4.1 Conclusions:

From the results data provided above, a number of important conclusions can be seen in the retrospective futility analyses in both clinical trials. The most important conclusion is that both futility analysis methods were accurate at predicting the final analysis results (i.e. would the trial have sufficient evidence to reject the null hypothesis, or not), and the interim analyses were most accurate as more data became available. The retrospective futility analysis findings are not very reliable at the 25% time point due to the volatility of the data this early in the trial. This was reflected by the variability seen in the outcome parameter statistics, such as the wide range of hazard ratio results observed, as can be seen from the very wide confidence intervals. In this case, in both trials the conservative setting for the testing the alternative hypothesis at a low significance level method (Method A) produced more results that suggested stopping the trial than in the aggressive setting. These interim futility analyses, however, become more accurate as more data was acquired, and the results became much more consistent with the true final analysis results even in the 50% analysis. At 75%, both methods were quite accurate in both trials. Upon reviewing the accumulative results of the retrospective interim analyses of both trials, however, one could surmise that a DSMC might have considered closing the OV.16 trial early for reasons of futility (a trial that appears to have shown a lack of difference in the treatment arms), had they been able to review the data at that time. One would also conclude that a DSMC would not have had reason to close the CO.17 trial early for findings of lack of efficacy, given the cumulative information provided from the interim analyses. In fact, one could argue that they might have been tempted to close early for reasons of benefit given the data available. This shows that futility analyses can be an effective and important tool for assessing a clinical trials success (or failure) in the interim, and can be a useful predictor for trial outcome. Incorporating futility analyses into the statistical analysis plans of clinical trials in a phase III setting, therefore, is beneficial.

Another important finding from this investigation is that, despite the increasing accuracy of both methods as more data became available, there still remains variability in the findings between the two methods. This variability is most apparent in OV.16, where significant differences persisted past the 25% analysis. This variability was seen in both the retrospective analyses of the original trial data, as well as with the cumulative

bootstrap sample analyses. Discrepancies between futility analysis method results could be seen in the overall cumulative final results of the bootstrap samples, but also within individual bootstrap samples as well. If one method would conclude that the trial should stop based on the results of that particular interim analysis, there was no guarantee that the other method would also reach the same conclusion with the same data. Although there was little variability in the CO.17 trial, a trial that did in fact find a significant difference in favour of the experimental arm, differences were still observed at the 25% analysis, as well as in the 50% analysis (but to a lesser extent). This was mostly due to the fact that this trial was positive, and therefore the probability of having a repeated sample by bootstrapping reach the threshold for futility was very small. The extent of the variability of retrospective futility analysis results observed with the OV.16 data, using both the original data set and the bootstrap samples, however, was a very interesting finding. These results were interesting because it had been hypothesized that there would not be much variability in the futility analysis method results within each trial, and, therefore, these results were unexpected. Instead, it could be seen that the stochastic curtailment method was significantly better at predicting the final analysis results of the trial than the other method, and these differences were even more apparent as more data became included in the analysis (and also, of course, as the investigation was increased when bootstrapping samples were analyzed).

This shows that although futility analysis is an important tool for monitoring clinical trial objectives, that perhaps multiple methods should be used in order to accumulate sufficient evidence to support decisions to continue or a stop a trial upon interim analysis, and certainly stochastic curtailment should be included in these assessments. A DSMC should be provided with as much evidence as is feasible in order to appropriately inform such decisions. We can see from the results provided at the interim analyses, that the DSMC would have likely had enough evidence to support closing the OV.16 trial early for reasons of futility, and would not have had enough evidence to do so for CO.17.

A third finding for this study was that there were significant differences when using different stopping rules for a futility analysis method. The aggressive setting produced significantly more decisions to stop the trial (for OV.16) than in the conservative setting, and this was seen for both futility analysis methods. This suggests that the stopping rules used for futility analysis can be significant, and should be considered when designing the statistical analysis plan of a clinical trial. This is especially true for interim analyses planned earlier in the course

of a clinical trial. It could be prudent to adjust these stopping rules accordingly, depending on the nature of the clinical trial or investigational agents. This will be discussed further below.

A final point of interest in this study is that the averaged results provided by repeated analysis (i.e. hazard ratio and confidence intervals, p-value, etc.) were comparable to the actual results obtained from the original data sets for each trial. The results of these repeated samples at each interim analysis also provide an interesting summary of a trials progress during the course of the study with respect to the primary endpoints. The cumulative results of futility analysis findings with the bootstrap samples (i.e. the decisions to continue or close the trial based using either Method A or B) also were consistent with the findings of the original data sets, in that both settings showed the stochastic curtailment method was a better predictor at determining trial outcome in this particular investigation.

In summary, futility analysis can be an effective tool at predicting the outcome of a clinical trial. This is most effective when using combined methods in order to provide sufficient evidence to support a decision, as the results between methods vary. Stochastic curtailment is a much more accurate predictor of trial outcome, when compared to testing the alternative hypothesis at a low significance level. The stopping rule guidelines used for planning futility analysis should also be considered, as the results produced from aggressive and conservative stopping rules differed significantly for both futility analysis methods examined. Finally, the statistics provided by bootstrapping provide an interesting picture of a completed clinical trials progress, appeared to be successful at increasing the robustness of the investigation, and were similar and consistent with the original trial data sets final analysis results.

4.2 Discussion:

In summary, this investigation aimed to evaluate the methodological and statistical principles associated with conducting analyses of futility by first performing a systematic review of the literature to determine trends in futility analysis methodology, and then applying these findings in retrospective futility analyses of two NCIC CTG trials that have reached final analysis. In doing so, the purpose was to provide insight into the effectiveness and accuracy of futility analysis methods, to promote consistency in clinical practice with interim analysis

planning, and to and to provide a basis for hypotheses-testing of optimum methodologies and their associated trade-offs. The overall objective of the study was to improve understandings of design, conduct and analysis of randomized controlled trials. The systematic review demonstrated that although there is a considerable amount of underreporting of interim analysis methodology, the two most common futility analysis methods observed were testing the alternative hypothesis at a low significance level and stochastic curtailment. These were used most frequently, but other methods were identified as well, and often combinations of methods were reported in the literature. The review also showed that there was variability in the use of aggressive and conservative stopping rules for these futility analysis methods. The retrospective futility analysis demonstrated that both methods can be an effective tool at predicting the outcome of a clinical trial, but there are inconsistencies in the results. The important findings of the retrospective futility analysis were that the stochastic curtailment method was most accurate at predicting the results of final analysis, that it may be beneficial to use more than one method when assessing futility at the interim in order to obtain more evidence to support a decision to stop or continue a trial, and that the results of futility analysis would differ significantly when using aggressive versus conservative stopping rules.

These results provided substantial information for addressing the objectives of this investigation, but there are some limitations to these results that require disclosure. The most obvious limitation, as previously addressed in describing the potential confounders of the systematic review, is the reporting bias observed in the literature that was associated with investigating interim analysis methods. These have been described in section 2.3.6, and the investigation was designed and undertaken in an effort to effectively limit this bias. Briefly, the issue is that there is considerable underreporting of futility analysis methods in the literature. This was especially apparent when trying to investigate the futility analysis methods used in trials where no significant findings were identified at interim analysis. To complicate this limitation further, this particular systematic review could not be performed by traditional means or guidelines, as there was not simply one exposure and outcome to investigate. Instead, this review investigated numerous methodologies used for interim futility analyses of randomized phase III clinical trials in oncology (i.e. numerous ‘exposures’), which were not associated with one particular outcome. Because of this, novel approaches to the execution of the systematic review had to be developed. The methods used were based on methods described in previous reviews identified

in the comprehensive/literary review undertaken at the start of this investigation. Validation methods were also incorporated as well, in order to confirm the accuracy of the methods used, and the services of Queens Library staff were also obtained to ensure the effectiveness of the literature search parameters.

The retrospective futility analysis findings were also not without limitations. For the purposes of this investigation, retrospective futility analyses were performed on two NCIC CTG clinical trials. These trials had previously published their primary analysis findings, and did not include an interim analysis in their original statistical plans. This exercise was very informative, produced valuable information, and was able to provide evidence to address the stated objectives. It would be interesting, however, to determine if these same conclusions would be reproduced if additional NCIC CTG clinical trials that had published final analysis results were included. This would be especially true if the additional trials included were similar to OV.16, and were also not able to provide sufficient evidence to reject the null hypothesis. If it had been feasible to expand the investigation and apply retrospective futility analyses to more clinical trials, and increase the original samples size from two trials, it could have produced a more compelling argument. Instead, bootstrapping methods were employed to increase the robustness of the investigation. The inclusion of additional clinical trials for a more complete investigation of futility analysis methods will be addressed further when discussing future considerations.

Despite these limitations, however, a number of interesting and important findings can be extracted from this investigation, and there are a number of implications for future interim analysis planning of clinical trials in a randomized phase III oncology setting. The results clearly demonstrate that futility analyses can be effective predictors of trial outcome, and especially at the 50% and 75% time points. Futility analyses, therefore, are effective tools for assessing the eventual success or failure of an RCT for meeting its objectives, and that the findings of futility analysis become more accurate as more information is accumulated and the data is less volatile. This is an important finding for statistical analysis planning of clinical trials when increased interim monitoring is warranted. This would be especially true for clinical trials investigating agents that are potentially quite toxic for patients, or if there is a concern for lack of efficacy compared to a standard of care treatment that has been proven to be effective. To a lesser extent, it is also important if the sponsor is concerned for the resources allocated to a clinical trial, as is increasingly the case in the current environment.

Although both methods were effective at predicting trial outcome, it is also important to note that there was a considerable amount of variability in the results obtained from these futility analyses. When one method provided evidence to support early termination, it was possible that the other method did not provide this same recommendation. In fact, it was clear that one method, stochastic curtailment, appeared superior to the other method at predicting trial outcome in this investigation. This demonstrated that in order to reach an informed decision as to whether a trial should continue, it may be beneficial to include multiple methods of assessing futility in the interim statistical analysis plans in order to gain better insight into the available data, and more evidence to support the decisions. Of course, stochastic curtailment should most certainly be included as a component of these multiple methods. It is interesting to see from the systematic review that a conditional probability was reported in 8 additional articles whose primary methodology was not stochastic curtailment. It can also be seen that 3 articles reported did not report a primary futility analysis method, but mention multiple methods used. This data shows us that there is considerable variation in the methods used for futility analysis, and that a number of trials do include additional methods to support a result from a primary method. This is an important trend for future clinical trials, and, as per the results of this investigation, appears to be an appropriate approach.

Finally, it is important to note that the stopping rules used will affect the recommendations from a futility analysis. This seems intuitive that aggressive stopping rules will result in a decision to stop more often than a conservative rule, but it was not expected that this difference would be so significant. This is also an important finding for planning future interim analyses. The stopping rules used should be considered when designing the statistical analysis plans of clinical trials, and there are instances where these rules may be more appropriate. For example, conservative stropping rules may be more beneficial when comparing an investigational agent that may not be much more efficacious than the standard, but has less associated toxicity. It may also be appropriate to use conservative rules if there is a known delayed treatment effect, or there is previous evidence of benefit for the investigational therapy. Conversely, aggressive methods may be more appropriate when the investigational agent is quite toxic, the standard is well established, or perhaps if the expected accrual rate is predicted to be considerably high.

The findings of this investigation introduce a number of interesting considerations and directions to guide future research on this topic. As previously mentioned, it would be beneficial to further the investigation of futility analysis method accuracy by expanding the retrospective futility analysis application to more clinical trials. Specifically, it would be interesting to perform this analysis on all available NCIC CTG clinical trials that meet the inclusion criteria, and pooling the results. This would make a more robust comparison of the two methods. Additionally, it would also be interesting to introduce additional methods that had been identified from the systematic review, and include those in the comparisons of the pooled analysis. Investigating the reliability of additional time points might be of interest as well, as some interim analyses are scheduled at approximately 30% and 60% of acquired total events.

An alternative direction to take future research would be to model futility analysis methods in specific settings relative to the primary endpoints of a trial. For instance, it would be interesting to investigate if a specific type is more adept at predicting trial outcome in DCR trials (documented clinical response), versus a time to event type trial (e.g. overall survival or progression/disease free survival). Aggressive and conservative stopping rules would also need to be included. Another interesting application would be to investigate if a particular futility analysis method would be more appropriate in trials that had a very fast (or very slow) accrual rate. There are a number of ways that research on futility analysis methodology can be advanced further, but the findings of this project have contributed to this research area.

In conclusion, it has been shown that futility analysis methods vary in the literature, and there are no specific guidelines for how to design and implement futility analysis testing in clinical trials. These interim analyses are important because their results can have serious implications for patient safety and clinical practice. The results of this study have provided a considerable amount of information on the topic of futility analysis methodology, and have helped to answer the question of how best to design and plan these analyses. This study has provided substantial information for guiding the practice of proper interim statistical analysis planning, and has therefore improved understanding of the design, conduct, and analysis of randomized phase III clinical trials in oncology.

REFERENCES:

1. (Authorship not provided): Draft guidance for clinical trial sponsors on the establishment and operation of clinical trial data monitoring committees. Federal Register 2001; 66:58151-58153.
2. Bassler D, Briel M, Montori V, et al. Stopping randomized trials early for benefit and estimation of treatment effects. JAMA 2010; 303(12): 1180-1187.
3. Bassler D, Montori M, Briel M, Glasziou P, Guyatt G: Early stopping of randomized clinical trials for overt efficacy is problematic. JCO 2008; 61:241-246.
4. Blay J-Y, Le Cesne A, et al. Prospective multicentric randomized phase III study of imatinib in patients with advanced gastrointestinal stromal tumors comparing interruption versus continuation of treatment beyond 1 year: The French sarcoma group. JCO 2007; 25(9): 1107-1113.
5. Booth C, Cescon D, Wang L, Tannock I, Krzyzanowska M. Evolution of the randomized controlled clinical trial in oncology over three decades. JCO 2008; 26(33):5458-5464.
6. Booth C, Ohorodnyk P, Zhu L, Tu D, Meyer R. Randomised controlled trials in oncology closed early for benefit: Trends in methodology, results and interpretation. European Journal of Cancer 2011; 47:854-863.
7. Bradbury P, Tu D, Seymour L, et al. Economic Analysis: Randomized Placebo Controlled Clinical Trial of Erlotinib in Advanced Non-Small Cell Lung Cancer. JNCI 2010;102(5):298-306.
8. De Roock W, Jonker DJ, et al: Association of KRAS p.G13D Mutation with Outcome in Patients with Chemotherapy-refractory Metastatic Colorectal Cancer Treated with Cetuximab. JAMA 2010 Oct 27; 304(16):1812-20.
9. Efron B, Tibshirani R: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science 1986;1(1):54-77.
10. Efron B, Tibshirani R: An Introduction to the Bootstrap. New York, NY, Chapman & Hall, 1993.
11. Emerson SS, Kittelson JM, Gillen DL. On the use of stochastic curtailment in group sequential clinical trials. UW Biostatistics Working Paper Series. Working Paper 243, *2005.*
12. Fleming, T. R., Harrington, D. P., and O'Brien, P. C. (1984). Designs for group sequential tests. Control Clin.Trials, 5(4):348—361
13. Fleming T, O'Brien: A multiple testing procedure for clinical trials. Biometrics 1979; 35:549-556.
14. Floriani I, Rotmensz N, et al. Approaches to interim analysis of cancer randomized clinical trials with time to event endpoints: A survey from the Italian National Monitoring Centre for Clinical Trials. Trials 2008; 9:46.
15. Friedlin B, Korn E. A comment on futility monitoring. Controlled Clinical Trials 2002; 23: 355- 366.
16. Freidlin B, Korn E. Monitoring for lack of benefit: A critical component of a randomized clinical trial. JCO 2009; 27(4): 629-633.
17. Geller NL: Planned interim analysis and its role in cancer clinical trials. J Clin Oncol 5:1485-1490, 1987.
18. Goldman B, LeBlanc M, Crowley J. Interim futility analysis with intermediate endpoints. Clin Trials 2008; 5(1):14-22.
19. Goodman SN: Stopping at nothing? Some dilemmas of data monitoring in clinical trials. Ann Intern Med 2007;146:882-887.

20. Goss P, et al. Randomized Trial of Letrozole Following Tamoxifen as Extended Adjuvant Therapy in Receptor-Positive Breast Cancer: Updated Findings from NCIC CTG MA.17. JNCI 2005; 97(17):1262-1271.

21. Grant A: Stopping clinical trials early. BMJ 2004;329:525-526.

22. Green SJ, Fleming TR, O'Fallon JR: Policies for study monitoring and interim reporting of results. J Clin Oncol 5:1477-1484, 1987.

23. Hoskins P, et al: Advanced Ovarian Cancer: Phase III Randomized Study of Sequential Cisplatin-Topotecan and Carboplatin-Paclitaxel vs Carboplatin-Paclitaxel. J Natl Cancer Inst 2010;102:1547-1556.

24. Jennison C, Turnbull BW: Group Sequential Methods with Applications to Clinical Trials. Boca Raton, FL, Chapman and Hall/CRC, 2000.

25. Jonker DJ, O'Callaghan CJ, et al: Cetuximab for the Treatment of Colorectal Cancer. N Engl J Med 2007; 357:2040-8.

26. Karapetis CS, Khambata-Ford S, et al. K-RAS Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer. N Engl J Med 2008 Oct 23; 359(17):1757-65.

27. Lachin J. A review of methods for futility stopping based on conditional power. Stat Med 2005; 24(18):2747-2764.

28. Lachin J. Futility Interim monitoring with control of type I and II error probabilities using the interim z-value or confidence limit. Clin Trials 2009; 6(6):565-573.

29. Lan K, DeMets D: Discrete sequential boundaries for clinical trials. Biometrika 1983; 70:659-663.

30. McNeil C. Negative data from lung cancer trial may change practical guidelines, study designs. JNCI 2006; 98(21):1518-1520.

31. Mittman N, Au H-J, Tu D, et al: Prospective Cost-Effectiveness Analysis of Cetuximab in Metastatic Colorectal Cancer: Evaluation of NCIC CTG CO.17 Trial. JNCI 2009;101(17):1182-1192.

32. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, et al: Randomized trials stopped early for benefit: a systemic review. JAMA 2005;294:2203-9.

33. Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH: Ethical issues in stopping randomized trials early because of apparent benefit. Ann Intern Med 2007;146:878-881.

34. Ng R, Hasan B, Mittman N, et al: Economic Analysis of NCIC CTG JBR.10: A Randomized Trial of Adjuvant Vinorelbine plus Cisplatin compared with Observation in Early Stage Non-Small Cell Lung Cancer – A Report of the Working Group on Economic Analysis, and the Lung Disease Site Group, NCIC CTG. JCO 2007;25(16):2256-2261.

35. Pampallona, S. and Tsiatis, A: Group sequential designs for one and two sided hypothesis testing with provision for early stopping in favour of the null hypothesis. Journal of Statistical Planning and Inference 1994; 42:19-35.

36. Skovlund E. Repeated Significance Tests on Accumulating Survival Data. J Clin Epi 1999; 52(11):1083-1088.

37. Snapinn S, Chen M, Jiang Q, Koutsoukos T. Assessment of futility in clinical trials. Pharm Stat 2006; 5(4):273-281.

38. Sparano JA, Fisher RI, et al. Randomized phase III trial of treatment with high-dose interleukin-2 either alone or in combination with interferon alfa-2a in patients with advanced melanoma. JCO 1993; 11(10): 1969-1977.

39. Strauss G, Herndon M, Maddaus D, et al. Randomized clinical trial of adjuvant chemotherapy with paclitaxel and carboplatin following resection in stage IB non-small cell lung cancer (NSCLC): Report of Cancer and Leukemia Group B (CALGB) Protocol 9633. JCO 2004; 22(14S).

40. Trotta F, Apolone G, Garattini S, Tafuri G: Stopping a trial early in oncology: for patients or for industry? BMJ 2004;329:525-526.

41. Ware J, Muller H, Braunwald E: The Futility Index, an approach to cost-effective termination of randomized clinical trials. Am J Med 1985; 78(4):635-643.

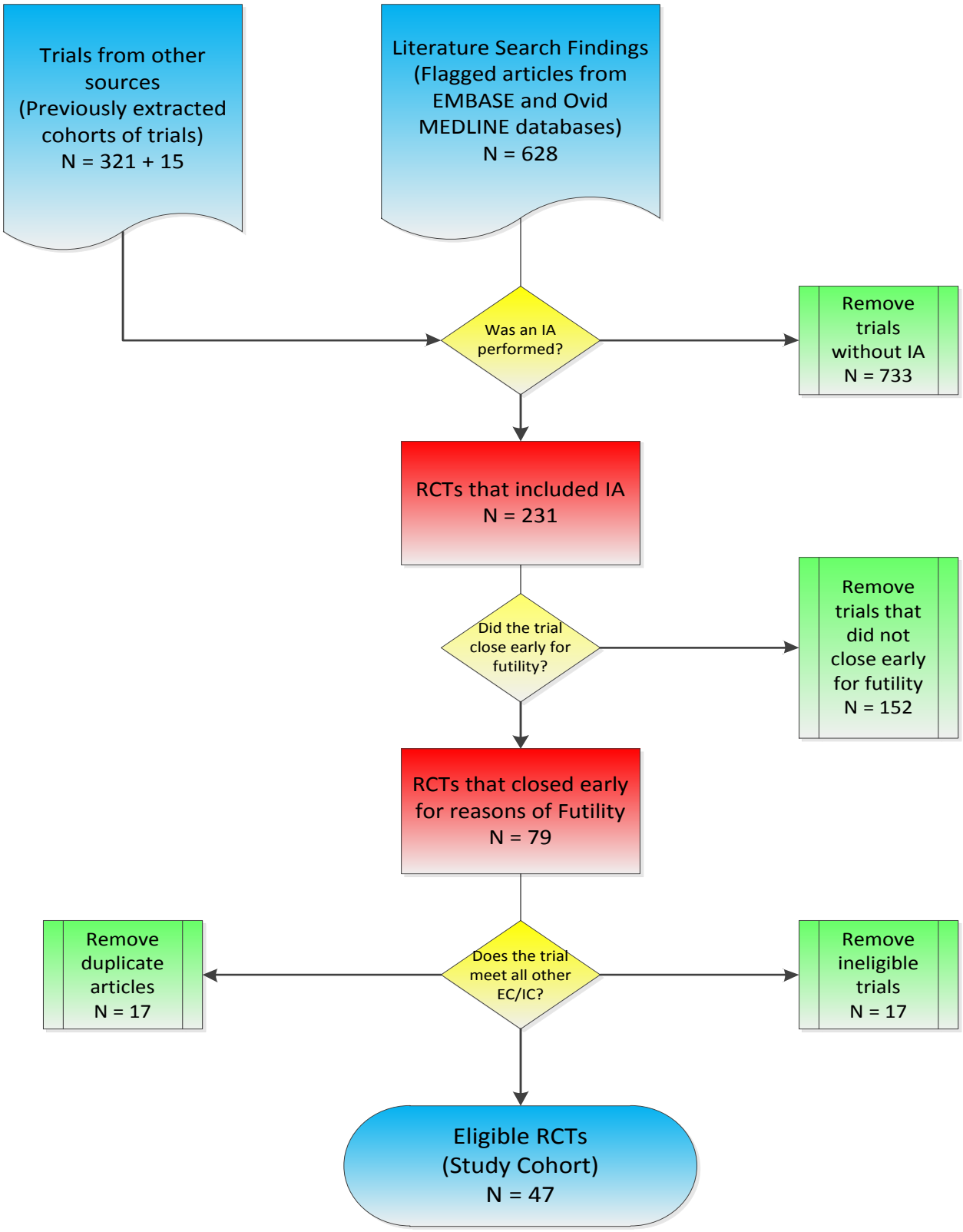
42. Webb P, Bain C, Pirozzo S. Essential Epidemiology: An introduction for students and health professionals. 2005. (Chapter 6)

43. Wheatley K, Clayton B: Be skeptical about unexpected large apparent treatment effects: The case of MRC AML12 randomization. Control Clin Trials 2003; 24:66-70.

44. Zhang Y, Clarke W. A flexible futility monitoring method with time-varying conditional power boundary. Clin Trials 2010; 7(3):209-218.

APPENDIX I:

Consort Flow Chart of Eligible Studies – Systematic Review:



APPENDIX II:

Futility Analysis Methodology – Systematic Review:

Data Abstraction Tool

Study Origin _____

A. STUDY METHODOLOGY

1. Disease site:
   ☐ breast ☐ lung ☐ gastrointestinal ☐ genito-urinary ☐ gynecological ☐ head & neck
   ☐ hematologic ☐ melanoma ☐ sarcoma ☐ supportive care/symptom control ☐ brain ☐ other

2. Setting:
   Non-metastatic: ☐ adjuvant ☐ neoadjuvant ☐ palliative
   Metastatic: ☐ first line ☐ > 1st line
   Hematologic: ☐ limited stage ☐ 1st line advanced ☐ >1st line advanced

3. Primary end-point for primary analysis:
   ☐ overall survival ☐ treatment response
   ☐ disease control (time to tx failure) ☐ companion reported outcome

4. If companion reported outcome:
   ☐ cost/economic analysis ☐ quality of life ☐ correlative biology ☐ symptom control ☐ other

5. If disease control:
   ☐ event free survival ☐ progression free survival ☐ Time to treatment failure ☐ other

6. Primary endpoint:
   ☐ single endpoint ☐ multiple/combined endpoints

7. Control group:
   ☐ no active treatment (placebo/no tx) ☐ active treatment (SOC, new tx)

8. Blinding:
   ☐ open ☐ single ☐ double

9. Number of arms:
   ☐ 2 ☐ 3 ☐ 4 ☐ >4

10. Intervention:
    ☐ radiotherapy ☐ surgery ☐ systemic ☐ supportive care/symptom control ☐ other

11. Equivalence study:
    ☐ superiority trial ☐ non-inferiority trial ☐ other

B. PROTOCOL DESIGN STATISTICS:

1. Hazard Ratio: _____ (or NA – not given)

2. Absolute Benefit: _____ (or NA – not given)

3. Sample Size:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

4. Power:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

5. Alpha:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

6. Early Termination?:
   ☐ yes ☐ no

7. Timelines for accrual:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

8. Timelines for follow up:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

9. Accrual rate:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

10. Was a prospective pre-defined statement/definition of futility analysis?
    ☐ ○ yes ☐ no ☐ amendment ☐ retrospective futility ☐ post interim analysis

12. If yes, was it done per definition?
    ☐ yes ☐ no ☐ NA

13. ITT (intent to treat) Analysis?:
    ☐ yes ☐ no

C. STUDY RESULTS

1. Futility Methodology:
   ☐ The power family of two-sided wedge tests [3]
   ☐ Error spending method [2]
   ☐ Stochastic Curtailment method based on conditional power (futility index) [4]
   ☐ Testing the alternative hypothesis at very low significance level. [1]

   References for above:

   [1] Fleming, T. R., Harrington, D. P., and O'Brien, P. C. (1984). Designs for group sequential tests. Control Clin.Trials, 5(4):348--361.
   [2] Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70:659--663.
   [3] Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one and two sided hypothesis testing with provision for early stopping in favour of the null hypothesis. Journal of Statistical Planning and Inference, 42:19--35.
   [4] Ware, J. H., Muller, J. E., and Braunwald, E. (1985). The futility index, an approach to the cost-effective termination of randomized clinical trials. Am J Med, 78(4):635--643.

2. Was a conditional power reported?:
   ☐ yes ☐ no

3. If no, can it be obtained?
   ☐ yes ☐ no

4. If yes (to either 2 or 3), Range of conditional power:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

5. Criteria for analysis:
   ☐ Number of events ☐ Number of patients

6. Number of events for final analysis:
   Planned _____ (or NA – not given)
   Actual _____ (or NA – not given)

7. Number of interim analyses:
   ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

8. Number of events for interim analysis:
   First: planned: _____ actual: _____
   Second: planned: _____ actual: _____
   Third: planned: _____ actual: _____
   Fourth: planned: _____ actual: _____
   Fifth: planned: _____ actual: _____

9. Did the interim analysis include an analysis of futility?:
   First: ☐ yes ☐ no ☐ unknown ☐ NA
   Second: ☐ yes ☐ no ☐ unknown ☐ NA
   Third: ☐ yes ☐ no ☐ unknown ☐ NA
   Fourth: ☐ yes ☐ no ☐ unknown ☐ NA
   Fifth: ☐ yes ☐ no ☐ unknown ☐ NA

10. Number of patients at futility analysis:
    First: planned: _____ actual: _____
    Second: planned: _____ actual: _____
    Third: planned: _____ actual: _____
    Fourth: planned: _____ actual: _____
    Fifth: planned: _____ actual: _____

11. Was the futility analysis done after all patients had been accrued?
    First: ☐ yes ☐ no ☐ unknown ☐ NA
    Second: ☐ yes ☐ no ☐ unknown ☐ NA
    Third: ☐ yes ☐ no ☐ unknown ☐ NA
    Fourth: ☐ yes ☐ no ☐ unknown ☐ NA
    Fifth: ☐ yes ☐ no ☐ unknown ☐ NA

12. Was the futility analysis done after all patients had completed treatment?
    First: ☐ yes ☐ no ☐ unknown ☐ NA
    Second: ☐ yes ☐ no ☐ unknown ☐ NA
    Third: ☐ yes ☐ no ☐ unknown ☐ NA
    Fourth: ☐ yes ☐ no ☐ unknown ☐ NA
    Fifth: ☐ yes ☐ no ☐ unknown ☐ NA

13. Based on their methods, was the futility considered aggressive or conservative?
    First: ☐ aggressive ☐ conservative ☐ NA
    Second: ☐ aggressive ☐ conservative ☐ NA
    Third: ☐ aggressive ☐ conservative ☐ NA
    Fourth: ☐ aggressive ☐ conservative ☐ NA
    Fifth: ☐ aggressive ☐ conservative ☐ NA

Based on Friedlin and Korn **Error! Bookmark not defined.**, there are five groups of futility boundary,

1. Moderately aggressive early stopping: with less 50% of information, reject the alternative hypothesis at 0.001 level
2. Aggressive early stopping: with less 50% of information, reject the alternative hypothesis at 0.005 level
3. Moderately aggressive late stopping: with 50% or more of information, reject the alternative hypothesis at 0.01 level
4. Aggressive late stopping: with 50% or more of information, reject the alternative hypothesis at 0.02 level
5. Conservative stopping: continue unless hazard ratio of new over the standard is greater than 1.

14. Action/results of Analysis:
    First: ☐ trial stopped for futility ☐ trial stopped for other reason ☐ trial continued ☐ NA
    Second: ☐ trial stopped for futility ☐ trial stopped for other reason ☐ trial continued ☐ NA
    Third: ☐ trial stopped for futility ☐ trial stopped for other reason ☐ trial continued ☐ NA
    Fourth: ☐ trial stopped for futility ☐ trial stopped for other reason ☐ trial continued ☐ NA
    Fifth: ☐ trial stopped for futility ☐ trial stopped for other reason ☐ trial continued ☐ NA

15. Result of final analysis:
    ☐ Benefit ☐ No benefit ☐ Not done

D. DEMOGRAPHICS/NATURE OF TRIAL

1. Sponsor (lead):
   ☐ cooperative group ☐ primary investigator ☐ industry

2. Collaboration:
   ☐ yes ☐ no ☐ industry only

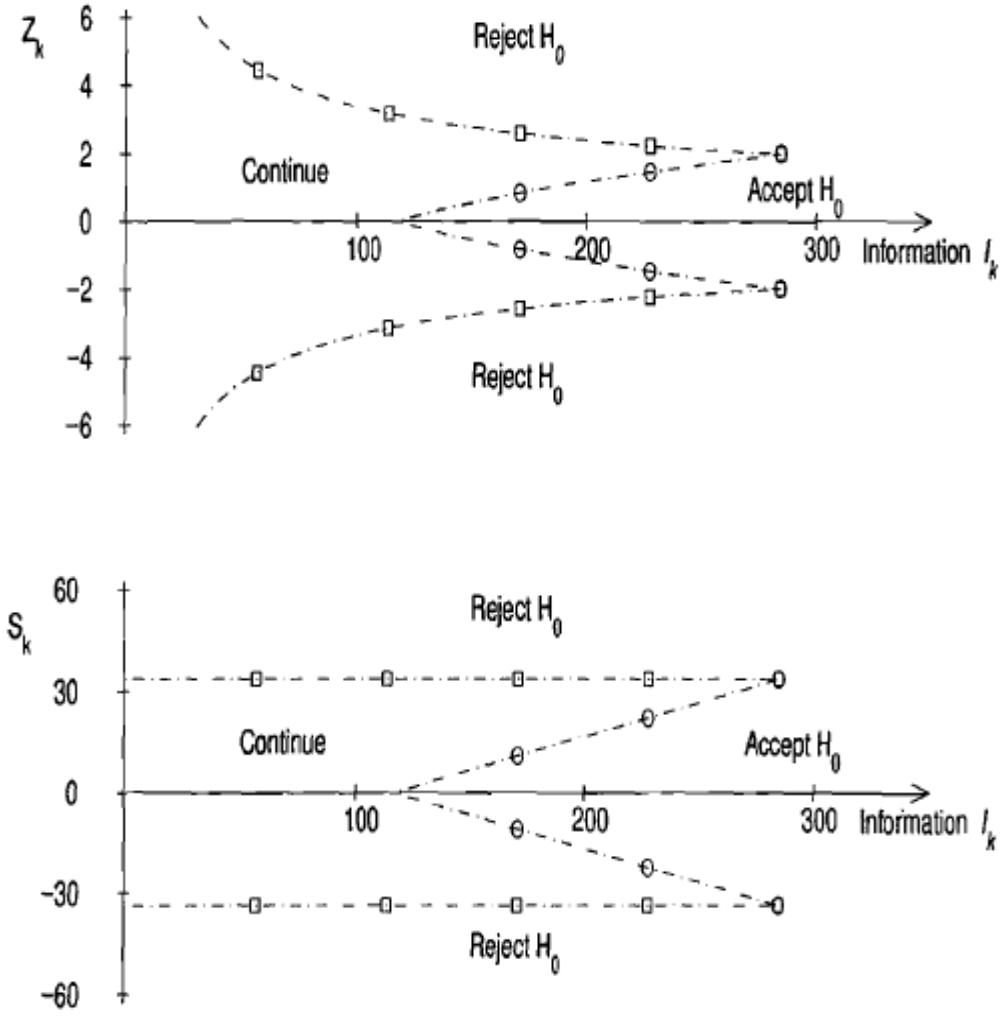
3. Participants/sites:
   ☐ single centre ☐ multicentre

4. International trial?
   ☐ yes ☐ no

APPENDIX III:

Example of Stopping Boundary Guidelines at Interim Analysis:

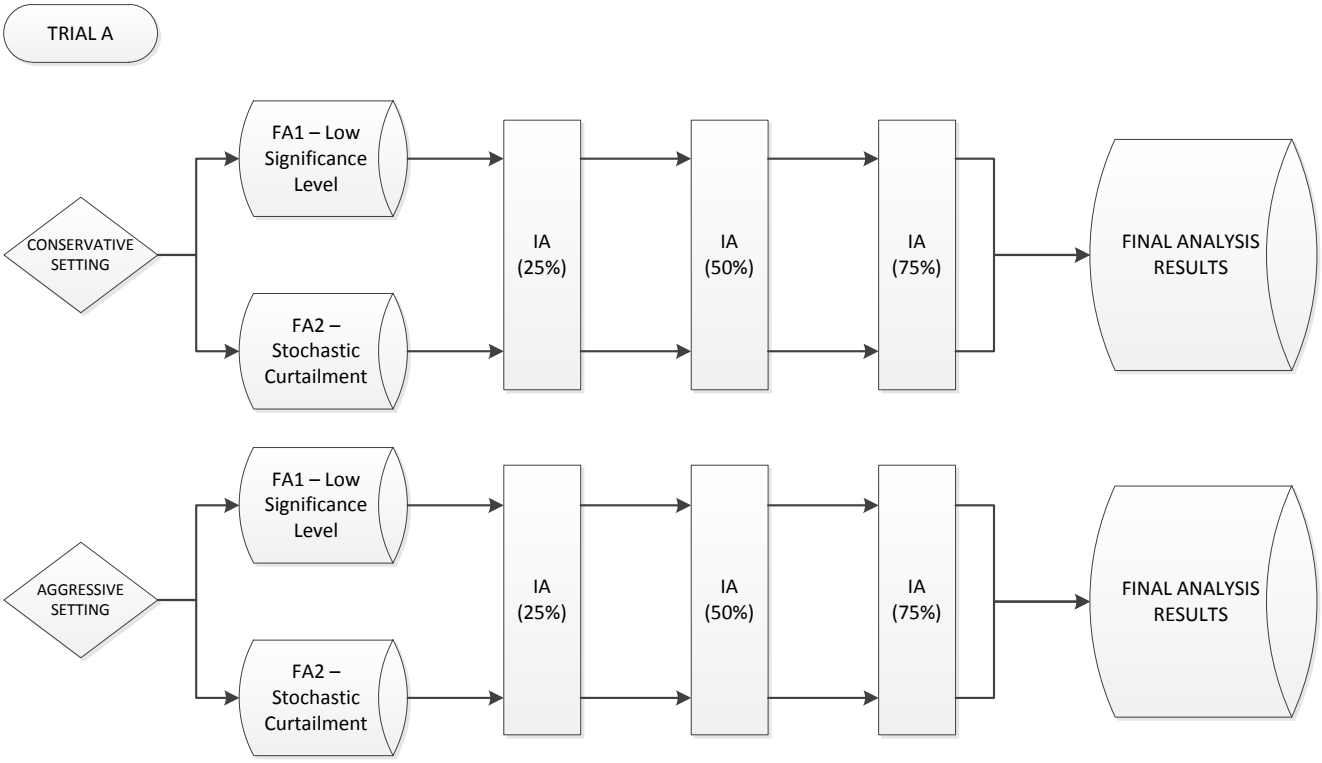
Figure 5.1 *A power family inner wedge test for five groups of observations*



Reference: Jennison C, Turnbull BW: Group Sequential Methods with Applications to Clinical Trials. Boca Raton, FL, Chapman and Hall/CRC, 2000.

APPENDIX IV:

Study Schema – Retrospective Futility Analysis:



APPENDIX V:

QUEEN'S UNIVERSITY HEALTH SCIENCES & AFFILIATED TEACHING HOSPITALS RESEARCH ETHICS BOARD-DELEGATED REVIEW

February 06, 2012

Mr. Chad Winch
Department of Community Health and Epidemiology
Queen's University

Dear Mr. Winch
Study Title: EPID-375-12 The Anatomy of Futility Analysis: An Investigation and Review of Futility Analysis Methods in Phase III Oncology Trials.
File # 6006575
Co-Investigators: Dr. R. Meyer, Dr. B. Chen

I am writing to acknowledge receipt of your recent ethics submission. We have examined the protocol (Thesis – Part A and Part B) for your project (as stated above) and consider it to be ethically acceptable. This approval is valid for one year from the date of the Chair's signature below. This approval will be reported to the Research Ethics Board. Please attend carefully to the following listing of ethics requirements you must fulfill over the course of your study:

Reporting of Amendments: If there are any changes to your study (e.g. consent, protocol, study procedures, etc.), you must submit an amendment to the Research Ethics Board for approval. Please use event form: HSREB Multi-Use Amendment/Full Board Renewal Form associated with your post review file **#6006575** in your Researcher Portal (https://eservices.queensu.ca/romeo_researcher/)

Reporting of Serious Adverse Events: Any unexpected serious adverse event occurring locally must be reported within 2 working days or earlier if required by the study sponsor. All other serious adverse events must be reported within 15 days after becoming aware of the information. Serious Adverse Event forms are located with your post-review file **6006575** in your Researcher Portal (https://eservices.queensu.ca/romeo_researcher/)

Reporting of Complaints: Any complaints made by participants or persons acting on behalf of participants must be reported to the Research Ethics Board within 7 days of becoming aware of the complaint. Note: All documents supplied to participants must have the contact information for the Research Ethics Board.

Annual Renewal: Prior to the expiration of your approval (which is one year from the date of the Chair's signature below), you will be reminded to submit your renewal form along with any new changes or amendments you wish to make to your study. If there have been no major changes to your protocol, your approval may be renewed for another year.

Yours sincerely,

Albert J. Clark.

Chair, Research Ethics Board
February 06, 2012

Investigators please note that if your trial is registered by the sponsor, you must take responsibility to ensure that the registration information is accurate and complete