STATISTICAL ANALYSIS PLAN

A DOUBLE-BLIND, RANDOMIZED, PHASE III TRIAL OF THE SAFETY AND EFFICACY OF CPP-1X/SULINDAC COMPARED WITH CPP-1X, SULINDAC AS SINGLE AGENTS IN PATIENTS WITH FAMILIAL ADENOMATOUS POLYPOSIS (FAP)

Protocol Number: CPP FAP-310

Statistical Analysis Plan

Version: Ver. 5.2, January 25, 2019

NCT Number: 01483144

Sponsor: Cancer Prevention Pharmaceuticals, Inc.

1760 E. River Road, Suite 250
Tucson, AZ 85718
Phone: 520-908-7774
Fax: 520-232-2191

STATISTICAL ANALYSIS PLAN (SAP) SIGNATURE SHEET

The undersigned have reviewed the format and content of this Statistical Analysis Plan (SAP) for Cancer Prevention Pharmaceuticals, Inc. study CPP FAP-310 titled "A DOUBLE-BLIND, RANDOMIZED, PHASE III TRIAL OF THE SAFETY AND EFFICACY OF CPP-1X/SULINDAC COMPARED WITH CPP-1X, SULINDAC AS SINGLE AGENTS IN PATIENTS WITH FAMILIAL ADENOMATOUS POLYPOSIS (FAP)" and have approved this SAP as finalized.

_____ _____

Bruce Levin, Ph.D. Date
Statistician

_____ _____

Robert B. MacArthur, Pharm.D., M.S. Date
Clinical Affairs and Development

_____ _____

Alfred M. Cohen, M.D. Date
Chief Medical Officer

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1. LIST OF ABBREVIATIONS

Abbreviation	Term
AE	Adverse Event
ANCOVA	Analysis of covariance
CPP	Cancer Prevention Pharmaceuticals, Inc.
CPP-1X	Eflornithine, DFMO, Difluoromethylornithine
CRO	Contract Research Organization
CTCAE	Common Terminology Criteria for Adverse Events
DMC/DSMB	Data Monitoring Committee/Data Safety Monitoring Board
EDC	Electronic Data Capture System
EFS	Event-Free Survival
EMA	European Medicines Agency
EORTC	European Organisation for Research and Treatment of Cancer
FAP	Familial Adenomatous Polyposis
FDA	Food and Drug Administration
FFQ	Food Frequency Questionnaire
HRQoL	Health Related Quality of Life
ICH	International Conference on Harmonization
ITT	Intent-To-Treat
LGI	Lower Gastrointestinal
LGIOI	LGI Observed Improvement
LOCF	Last Observation Carried Forward
NASR	Nutrition Assessment Shared Resource (Fred Hutchinson Cancer Research Center)
NCI	National Cancer Institute
ODC	Ornithine Decarboxylase
QLQ	Quality of Life Questionnaire
SAM	S-Adenosyl Methionine
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SAS	Statistical Analysis System
UGI	Upper Gastrointestinal
UGIOI	UGI Observed Improvement
Z-score	The value of the critical ratio (score statistic divided by its standard error) based on the stratified log-rank statistic.

2. INTRODUCTION

This Statistical Analysis Plan (SAP) describes the proposed statistical analysis of the Cancer Prevention Pharmaceuticals, Inc. (CPP) study CPP FAP-310 entitled: "A DOUBLE-BLIND, RANDOMIZED, PHASE III TRIAL OF THE SAFETY AND EFFICACY OF CPP-1X/SULINDAC COMPARED WITH CPP-1X, SULINDAC AS SINGLE AGENTS IN PATIENTS WITH FAMILIAL ADENOMATOUS POLYPOSIS (FAP)".

The purpose of this document is to apply sound analytic principles through description and pre-specification of the statistical approaches and data handling conventions for key analyses and the randomization processes.

This plan will focus on analysis of the primary and secondary endpoints, intended to assess the extent to which the combination treatment (CPP-1X 750 mg daily + sulindac 150 mg daily) is more effective than each of the two single agent treatments alone (CPP-1X 750 mg daily + placebo, and sulindac 150 mg daily + placebo) in delaying the time from the date of randomization to the date of the first occurrence of any FAP-related event in FAP patients. Sample size and total FAP-related events will be discussed, along with study population definitions. In addition, we will summarize our approach regarding the analysis of safety data, pharmacokinetic data, and patient reported quality of life data. The composition, charter, and initial tasks of the Data Monitoring Committee (DMC) will be described, along with planned interim analyses.

3. STUDY OBJECTIVES AND ENDPOINTS

3.1 Overview of Study

This is a randomized, double-blind, phase III trial in patients with FAP to evaluate the safety and efficacy of the combination product 750 mg CPP-1X + 150 mg sulindac versus each of the two single-agent products: 1) 750 mg CPP-1X; and 2) 150 mg sulindac. The primary study comparisons will be made between the combination product and each single agent product, separately.

Eligible patients who have given informed consent will enter the study with the intent to participate for the full treatment period of 24 months. Based on a subject's date of randomization, patients will be offered continued preparticipation in blinded treatment for up to a total of 36, 42 or 48 months of treatment if they have completed the initial 24 months of treatment without an FAP event until one of the following occurs: 1) subject has an FAP event or comes off study for other reasons, 2) all randomized subjects have reached a minimum of 24, 36, 42 or 48 months of treatment.

A stratified randomization procedure will be used based on FAP-related time-to-first-event prognosis. The event prognosis groups are represented by 1) best (i.e., longest projected time to first FAP-related event) - rectal/pouch polyposis, 2) intermediate - duodenal polyposis, and 3) worst - pre-colectomy. If a subject has two or more of these disease sites, the most severe prognosis stratum will be assigned for randomization (e.g. worst > intermediate > best). Since an individual may have more than one disease site involved, the trial will assess time to any defined FAP-related event in the patient as a whole. In order to minimize potential treatment arm imbalance a centralized randomization process will be used to balance treatment groups within disease prognostic strata.

The stratification is predicated on the following projected event rates for the disease site specific eligibility criteria (rectal/pouch polyposis, duodenal polyposis, pre-colectomy) which are provided in the clinical study protocol (Appendix E) and summarized in Table 3.1.

Table 3.1 Projected Event Free Rate by Disease Site

Disease Site	Projected 2-year Event Free Rate Without Treatment
Rectum/Pouch Polyposis	Excisional intervention and/or high risk adenoma – 40 - 60%
Duodenal Polyposis	Excisional intervention, Spigelman stage progression, cancer – 50%
Pre-colectomy	Colectomy, proctocolectomy – 10%

The stratified randomization will be reflected in stratified analyses for the primary and secondary endpoints. For design purposes only, however, the following simplified assumptions were made: an overall event free proportion of 30% in the single agent treatment group and 60% in the combination treatment group after 24 months of study treatment. This is described further in Table 3.2 below. See Section 5.1 for sample size and power calculations.

Table 3.2 Overall Event Free Proportions After Two Years of Follow-up*

Treatment	S(t)	t (months)	Median Time to Event (months)
Combination	0.6	24	32.566
Single Agent	0.3	24	13.817

*These proportions are assumed for design purposes only. The median times to event are based on the assumption of an exponential time-to-event function S(t) in each group.

At least 150 eligible patients will be enrolled in this study, with at least 50 per treatment group. Patients will be randomized to one of three treatment groups within the prognostic strata defined above in equal proportions (i.e., 1:1:1 randomization): 1) CPP-1X plus sulindac, 2) CPP-1X placebo plus sulindac, 3) CPP-1X plus sulindac placebo. The study is double blinded, so neither patients nor Investigator nor Sponsor will be aware of treatment assignment.

The grading of adverse events will be based on the current version of the NCI Common Terminology Criteria for Adverse Events (CTCAE Version 4.03).

Refer to the CPP FAP-310 protocol for the details of the treatment and study schedule, and description of study procedures and tests.

3.2 Study Objectives

The primary study analysis will be performed using the intent-to-treat (ITT) population which is defined in Section 4.1. Statistical methods for the primary analysis are described in Sections 5.1 and 5.5.3.

The primary objective of this trial is to determine whether the combination of CPP-1X + sulindac is superior to either single-agent treatment individually in delaying time from the date of randomization to the date to the first occurrence of any FAP-related event.

Thus the primary objective contains two treatment comparisons:

1. CPP-1X placebo + sulindac active vs. CPP-1X active + sulindac active,
   and
2. CPP-1X active + sulindac placebo vs. CPP-1X active + sulindac active

These two treatment comparisons will be performed sequentially, as described below.

The combination of CPP-1X active + sulindac active is specified as the reference treatment group because it is common to both comparisons. In addition, because the purpose of the combination treatment is to delay the time from randomization to FAP-related disease progression compared to single-agent treatments, formulating the hypothesis tests in this manner will allow a positive rather than a negative *Z*-score for the test statistic to be interpreted as supportive of this purpose.

Each comparison will be performed at the 2-sided 0.05 level of statistical significance.

As explained in Section 3.2.1 below, the decision to seek regulatory approval based upon the results of the primary objective will be taken sequentially.

3.2.1 Primary Analysis

The primary analysis will be a time-to-event analysis using the stratified log-rank test. Graphical analyses (log-minus-log plots) will be used to check the assumption of constant hazard ratios with the COX model. See Section 7.3 below for further details of the calculation of the stratified log-rank statistic. The strata are the patient's sites of disease involvement at baseline, which is determined prior to randomization, and are: Rectal/pouch polyposis, Duodenal polyposis, and Pre-colectomy.

For the primary analysis, two stratified log-rank tests will be performed with treatment coded as a binary value (i.e., 0 or 1). Time to event curves will be displayed using the method of Kaplan and Meier[8]. Additional analyses involving the overall 3-treatment group comparison and use of additional study populations (see Section 4.) for the two pairwise treatment comparisons, will be performed as supplemental analyses. Refer to section 5.5.3 for further details.

3.2.2 Regulatory Review of Primary Outcome Compliance

CPP has received advice from both the FDA and the EMA concerning the test method and criteria (Sections 3.2 and 5.) to be used when analyzing the primary study outcome, including the value of alpha.

The US Food and Drug Administration (FDA) noted that the comparison of single agent CPP-1X vs. combination treatment will not provide information about the treatment effect of CPP-1X (i.e. eflornithine), and therefore this result will not be included in the product label. Therefore, the primary analysis of import to FDA is the comparison of single agent sulindac vs. combination treatment.

The European Medicines Agency (EMA) noted that for the three arm trial design the superiority of the combination treatment must be tested against both mono-components. The combination treatment must be shown to be superior to both mono-components, or else the results will be considered inconclusive. As a consequence, consistent with EMA/SAWP commentary, alpha does not need to be adjusted for multiplicity and two separate 5% two-sided tests will suffice.

To harmonize the advice provided by these two Agencies, CPP will sequentially perform the two primary comparisons as part of the primary analysis, each at the 2-sided $p = 0.05$ level. All information concerning these comparisons will be clearly provided to both Agencies. The single treatment comparison requested by FDA will be available, as will the two comparisons requested by EMA, both at the requested level of alpha. This approach fulfills the differing requirements for the primary comparison as asked for by each Agency.

We note that this approach is both a fixed-sequence and gatekeeping approach. It is fixed-sequence in that the comparison of combination with single-agent Sulindac takes place before the comparison of combination with single-agent Eflornithine and the first serves as a gatekeeper for the second (i.e., no declaration of significance in the second comparison will be made if the first comparison is not significant at the 0.05 level). Therefore, the type I error in the sequential testing is well controlled. In addition, because *both* tests must be significant for EMA approval, the type I error of the second test in the sequence is less than 0.05.

3.2.3 Secondary Efficacy Analyses

Any improvement observed by the investigator during upper gastrointestinal (UGI) and lower gastrointestinal (LGI) visualization (i.e. endoscopy and colonoscopy) at the 6 and 12-month study visits will be described using the variables UGI Observed Improvement (UGIOI), and LGI Observed Improvement (LGIOI). Each patient will have one pair of UGIOI and LGIOI outcomes.

UGIOI and LGIOI are binary outcomes derived from numerical determinations (henceforth, "investigator change scores" or more briefly, "scores") assigned by the investigator during each procedure, using a scale (–2, –1, 0, +1, +2) which corresponds, respectively, to the investigator's overall qualitative assessment of: much worse, worse, no change, improved, much improved. At the month 6 procedures the investigator scores UGI and LGI findings as changes from baseline. At the month 12 procedures, the UGI and LGI findings are scored relative to the month 6 procedures.

The UGIOI (and respectively, the LGIOI) secondary endpoint independently summarizes the corresponding 6- and 12-month investigator change scores according to whether or not there was *any positive improvement* at either month 6 (compared to baseline) or at month 12 (compared to baseline or month 6), under the condition that there be *no worsening at either timepoint* (compared to the preceding timepoint). Here are the specific possibilities (where "Improvement" stands for either the UGIOI or LGIOI secondary endpoint):

- If the 6-month score is –2 or –1, Improvement=NO irrespective of the 12-month score.
- If the 6-month score is 0, then Improvement=YES if and only if the 12-month score is +1 or +2. Otherwise Improvement=NO.
- If the 6-month score is +1 or +2, then Improvement=YES if and only if the 12-month score is greater than or equal to 0. Otherwise Improvement=NO.

Any patient who drops out of the study before the month 12 assessment will be considered Improvement=NO.

These binary UGIOI and LGIOI secondary efficacy endpoints will be compared in a manner analogous to the primary analysis, using the same two primary treatment comparisons (1: CPP-1X placebo + sulindac active vs. CPP-1X active + sulindac active; and 2: CPP-1X active + sulindac placebo vs. CPP-1X active + sulindac active), and conditioning on the three disease site strata (Rectum/Pouch Polyposis, Duodenal Polyposis, Pre-colectomy). The null hypothesis of no association between treatment group and Improvement endpoints will be tested using the exact Mantel-Haenszel procedure for combining the evidence contained in the fourfold tables (Treatment=Single agent vs. Combination cross-classified by Improvement=YES vs. NO) across the three strata. For each of the two treatment comparisons, exact Mantel-Haenszel p-values will be calculated for both the UGI and LGI assessments (using the point-probability method based

on the convolution of three independent central hypergeometric distributions; see [1]).

The overall type I error for the secondary efficacy analysis will be controlled using the Hochberg step-up method for multiple comparisons[2]. This analysis will be performed in the ITT population. The primary analysis will serve as a gatekeeper to control the overall type I error rate at 0.05 for both primary and secondary analyses. That is, significance for the secondary efficacy analysis will be declared only if the primary p-value is 0.05 or less, when the p-values are tested sequentially per the Hochberg method[2].

Further analyses of the secondary endpoints may be conducted to evaluate and provide additional evidence to support its validity and confirm its clinical relevance.

3.2.4 Other Secondary Outcomes

To explore how study treatment group relates to other efficacy outcomes, genotype, phenotype, disease locations and endoscopic findings, additional analyses are planned. These analyses will be performed in the ITT group, the Per Protocol Group, and other defined subgroups (see Section 4 Populations for Analysis) wherever possible and will all be clearly noted as such.

The UGIOI and LGIOI outcomes will be tabulated and summarized using the month 6 visit scores, alone. Similarly, the UGIOI and LGIOI outcomes will tabulated and summarized across all study visits.

As both part of the primary analysis, and further explored in these additional analyses, median time to event for each treatment group will be determined. This will be explored for each of the study populations (i.e. ITT, per protocol, and others), study disease stratum groups, and in the Disease Site subgroups (see below).

Pharmacokinetic data (plasma concentrations measured at patient visits) will be used to estimate population pharmacokinetic parameters for the CPP-1X (eflornithine), sulindac, and CPP-1X (eflornithine) + sulindac treatment groups (i.e. for each analyte for those patients on combination treatment).

The subcategories of FAP events will be explored by disease stratum groups, and by Disease Site subgroups (see below). The subcategories of FAP events include:

Table 3.3 Disease Site Subgroups

No	Group	Description
1	Disease Progression Indicating Need for Colectomy with IRA or Total Procto-Colectomy	Applies only to subjects with FAP surgical status of precolectomy (field name: DIAGSXST, code 1)
2	Excisional intervention by surgical snare or trans-anal excision to remove any polyp >=10mm in size (per pathology report) and/or pathologic evidence of high grade dysplasia	Excisional intervention does not apply to subjects with FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4) or subjects with FAP surgical status of precolectomy (field name: DIAGSXST, code 1). For all remaining subjects, all >5 mm polyps must have been removed at baseline also

No	Group	Description
		High grade dysplasia does not apply to subjects with FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4)
3	Disease Progression Indicating Need for Proctectomy	Does not apply to subjects with FAP surgical status of proctocolectomy with ileal pouch anastomosis (IPAA) (field name: DIAGSXST, code 3) or FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4)
4	Disease Progression Indicating Need for Pouch resection	Does not apply to subjects with FAP surgical status of precolectomy (field name: DIAGSXST, code 1). or FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4) or FAP surgical status of colectomy with ileorectal anastomosis (IRA) (field name: DIAGSXST, code 2)
5	Development of cancer in rectum or pouch	Does not apply to subjects with FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4)
6	Progression in Spigelman Stage to more advanced stage (Stage 2, 3, 4)	Does not apply to subjects without duodenum (determined by abdominal surgical procedure of duodenectomy, duodenohemipancreatomy, or Whipple procedure (field name SXPROC) and screening/baseline upper GI endoscopy not done (field name UGIND, Boolean data type)
7	Disease Progression indicating need for excisional intervention (sub-mucosal resection, trans-duodenal excision, duodenectomy, ampullectomy, Whipple procedure)	Does not apply to subjects without duodenum (determined by abdominal surgical procedure of duodenectomy, duodenohemipancreatomy, or Whipple procedure (field name SXPROC) and screening/baseline upper GI endoscopy not done (field name UGIND, Boolean data type)
8	Development of cancer in duodenum	Does not apply to subjects without duodenum (determined by abdominal surgical procedure of duodenectomy, duodenohemipancreatomy, or Whipple procedure (field name SXPROC) and screening/baseline upper GI endoscopy not done (field name UGIND, Boolean data type)
9	Spigelman stage (no progression)	Does not apply to subjects without duodenum (determined by abdominal surgical procedure of duodenectomy, duodenohemipancreatomy, or Whipple procedure (field name SXPROC) and

No	Group	Description
		screening/baseline upper GI endoscopy not done (field name UGIND, Boolean data type)
10	>= 10 mm polyp in rectum pouch	Does not apply to subjects with FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4) or subjects with FAP surgical status of precolectomy (field name: DIAGSXST, code 1). For all remaining subjects, all >5 mm polyps must have been removed at baseline.
11	High grade dysplasia (rectum-pouch)	Does not apply to subjects with FAP surgical status of colectomy with ileostomy (field name: DIAGSXST, code 4)

These subcategories will be analyzed within the assigned disease strata (i.e. assigned at time of randomization). Separately, they will be analyzed by Disease Site, according to the concept of "the patient as a whole" meaning that *all known disease sites will be considered* because patients more often than not have more than one diseased organ system. For example, a patient randomized to treatment in the Duodenal Polyposis stratum who is also known to have colonic polyps will appear in the analyses for both Disease Site subgroups (Duodenal and Colonic). The total number of patients in the Duodenal Disease Site subgroup will include, for example, all patients randomized in the Duodenal Polyposis stratum + all other patients with known duodenal disease (irrespective of whether they have rectal/pouch disease). A similar approach will be taken for the Rectum/Pouch Polyposis group and Pre-colectomy strata.

The presence or absence of ODC polymorphisms, including the single nucleotide polymorphisms (SNPS) rs2302615 and rs2302616 and their relation to treatment group and outcome will be tested with the likelihood ratio test.

The excretion of 5 urinary polyamines (diacetylspermine, n1-acetylspermidine, n8-acetylspermidine, decarboxylated SAM, and putrescine) will be assessed in relation to treatment group and outcome, using the single point concentration data gathered from the urine samples harvested at each study visit.

Patient reported health related quality of life measures will be evaluated using HRQoL (see Section 5.6 Health Related Quality of Life (HRQoL) Statistical Methods for more details).

Tissue and dietary polyamine levels, as collected at patient study visits will be analyzed together with the results of the dietary questionnaires (see Section 5.7 Dietary Assessment) and related to treatment group and study outcomes.

Safety outcome data and analyses are described in Section 5.5.5 Assessment of Toxicities.

4. POPULATIONS FOR ANALYSIS

4.1 Intent-to-Treat (ITT) Population

The intent-to-treat population includes all patients that have been randomized to one of the three study arms. Patients will be analyzed in the group to which they were randomized, whether or

not they received their assigned treatment, any treatment whatsoever, or completed their treatment course and follow-up.

4.2 Safety Population

The safety population is defined as all ITT patients who received at least one dose of study medication. Patients who do not receive any study treatment (CPP-1X or sulindac or their combination) are excluded from this population. Patients will be analyzed in the treatment group according to which actual treatment was initially received.

4.3 Per Protocol Population

The per-protocol population is defined as the subset of the ITT population that fulfill all protocol eligibility, intervention, and outcome assessments.

4.4 Other Populations

Within the entire study patient population there will be subsets who did not receive the full course of per protocol treatment. The major indicators for premature withdrawal are delineated below. The patient diary and pill count will define the extent of treatment compliance during the study.

For exploratory and sensitivity analyses the following subsets will be included in secondary analyses:

- Patients withdrawn for personal reasons
- Treatment discontinued because of disease symptoms
- Treatment discontinued because of patient symptoms
- Compliance <80% treatments taken
- Treatment discontinued because of intercurrent medical or surgical illness.

5. DETERMINATION OF SAMPLE SIZE AND STATISTICAL METHODS

5.1 Determination of Sample Size

The primary endpoint of this trial, time to meaningful clinical events in an orphan disease population, is novel and to date there are no published trials to draw upon that have incorporated the exact FAP-related endpoint of this trial. Available data from primary literature sources include clinical studies where polyps were counted over a fixed time period, in different FAP populations (see protocol for tabulated listing).

From these data a reasonable range of event frequencies was estimated to produce the sample size and power calculations incorporated into this trial. These time-to-event estimates were reviewed by key FAP opinion leaders prior to finalization of the study design. The following reflects the possible range of FAP events it was thought plausible to observe.

5.2 Power Calculation Assumptions

1) The level of statistical significance is set at 0.05, using a 2-sided stratified log-rank test for time-to-first FAP-related event in continuous time, for each of the two between-group comparisons (i.e. single agent sulindac vs. CPP-1X plus sulindac and single agent CPP-1X vs. CPP-1X plus sulindac). The only covariates in the log-rank test will be the treatment groups (see section 7.3 for details);

2) A doubling of the two-year event-free proportion from either of the single agent treatment arms to the combination treatment group;

3) Power of at least 85% to detect the above-mentioned treatment effect comparing either of the two single treatment arms to the combination arm;

4) The two single-agent treatment groups have approximately the same event rate.

The following calculations are based on our review of limited single-agent data for eflornithine and sulindac, where FAP clinical trial primary endpoints involved polyp counting. Extrapolating these data to two-year event-free proportions implies a single overall two-year event-free proportion of at least 60% to 70% for the combination treatment group and 30% in each single agent treatment group. This is described further in Table 3.2.

5.3 Hazard Rates

Because the power of time-to-event analyses depends on the total number of observed primary endpoints ("events") and the hazard ratio in a given two-arm comparison of a single-agent versus combination therapy, we translate the above doubling of two-year event-free proportions into hazard ratios under a simplifying assumption of exponentially distributed time-to-event. Note that the primary analysis described below, in Section 7.3, does not rely on such an assumption and provides a statistically valid test of the null hypothesis even if hazard rates are not constant. Furthermore, the stratified log-rank test is an optimal test (locally most powerful) under the assumption that the *ratio* of the two groups' hazard functions remains constant over time (the proportional hazards assumption). Note that the much stronger assumption, that the individual hazard functions themselves remain constant over time, would be dubious in this trial. Therefore, irrespective of how the two-year event-free proportions are translated into hazard ratios, it is the latter which forms the *design alternative parameter* for the trial.

Under the exponential assumption, the hazard ratio (HR) comparing one treatment arm to another is given by the natural logarithm of the two-year event-free proportion for the first arm divided by the natural logarithm of the two-year event-free proportion for the other arm. Thus if the combination arm is assumed to have a two-year event-free proportion of 60%, which is double that of the 30% two-year event-free proportions assumed for the single-agent arms, the HR is $\{\log(0.60) / \log(0.30)\} = 0.4243$. *This is the design alternative hazard ratio for this trial* as it represents the minimum clinically meaningful treatment effect desired for the combination therapy compared to either single-agent therapy. Insofar as the combination therapy may have a two-year event-free proportion of *at least* 60%, and may prove to be perhaps 70% or greater, the design alternative HR of 0.4243 is conservative; the true (albeit unknown) HR is thought possibly to range from 0.4243 down to $0.30 = \{\log(0.70) / \log(0.30)\}$ (see discussion below).

Given that the primary hypotheses are stated in terms of comparing either single-agent arm to the combination arm, we note that the equivalent design alternative hazard ratio becomes $\{\log(0.30) / \log(0.60)\} = 1/0.4243 = 2.357$.

For the anticipated range of hazard ratios, 25 to 49 events would be needed for each two-group comparison at the 2-sided 0.05 level to achieve 85% power [3, 4]. Assuming two-year event proportions of 70% in either of the two single-agent groups and 30% to 40% in the combination arm with 50 patients per arm, the expected number of patients with an FAP-related event in either of the two single-agent groups would be 35 and 15 - 20 in the combination arm. The study design expectation is to have 50 - 55 patients with a FAP-related event in each two arm comparison, achieving at least 85% power under the design alternative. The standard deviation around the expectation of 55 events is 4.74, so observing the required number of 49 events or more would be highly likely (the probability is about 91%). If the total number of events in either comparison were only 43, there will still be 80% power to declare a significant treatment

difference under the design alternative of 0.4243.

As the two-year event proportion in the combination arm decreases from 40% with a corresponding decrease in the hazard ratio, the likelihood of observing the required number of events to maintain 85% power actually increases. For example, at the lower expectation of 50 events arising from an assumed two-year event proportion of 30% in the combination arm, the standard deviation of the total number of events in a two-arm comparison decreases to 4.58 and the probability that the observed number of events will exceed the 25 required to achieve 85% at a HR of 0.30 is virtually certain.

5.4 Data Monitoring Committee (DMC)DMC General Information

A Data Monitoring Committee (DMC) will oversee the performance and safety conduct of this study. The DMC will consist of at least three members (two MDs and one statistician as voting members) who will receive confidential reports on a periodic basis. The DMC will be responsible for decisions regarding possible termination of the study for either futility or safety reasons.

A detailed DMC Charter will be produced separately by the DMC membership. It is anticipated that any reviews of study data will be performed in a blinded manner, looking at pooled data (all treatment groups combined into one group) to assess mission-critical parameters such as overall recruitment and event rates. Of course, patient safety issues take precedence over bias-protection and control of type I error, and so the DMC will have the privilege of breaking the blind on a need-to-know basis if safety issues of concern arise in order to consider risk-benefit issues. Details concerning DMC responsibilities and duties may be submitted as a stand-alone document to the FDA and EMA, including items such as specification of early termination rules and other matters as the DMC deems to be important and relevant to the ethical conduct of this study.

CPP will inform the DMC that there will be two study evaluations for the DMC to consider during the trial, one interim look for sample size reassessment and one look for efficacy and futility.

The method for reassessment of sample size is based upon the FDA Guidance, *Adaptive Design Clinical Trials for Drugs and Biologics (February 2010)*. There will be no hypothesis testing. The DMC will assess the observed trial event rate based on pooled data only. They will make a recommendation to the sponsor on whether the pooled event rate is sufficient to preserve the integrity of the trial, and if not, to recommend a revised sample size. For this assessment the study statistician will, if possible, estimate the overall observed event rate and 90% confidence interval. This assessment will be performed using data from a single time point, when enrollment is approximately 95% complete. If this type of assessment is not possible, then an assessment will be performed taking into consideration the total number of subjects randomized, total number of events, total number of dropouts, and cumulative study safety data.

5.4.2 Prespecified Interim Efficacy and Futility Analysis

A pre-specified interim efficacy and futility analysis will be conducted as described below. The assessment will be performed after a total of 45 primary endpoints have occurred, which represents 50% of expected maximum trial information, or as soon thereafter as possible. After reviewing the analysis results in a closed session, the DMC will provide recommendations regarding possible termination of the study for either futility, efficacy, or safety reasons, to the Sponsor Steering Committee.

The analysis will be performed for each of the two treatment comparisons contained in the

primary objective:

1. CPP-1X placebo + sulindac active vs. CPP-1X active + sulindac active,

   and

2. CPP-1X active + sulindac placebo vs. CPP-1X active + sulindac active.

As stated above in Section 3.2, the combination of CPP-1X active + sulindac active is specified as the reference treatment group because it is common to both comparisons and formulating the hypothesis tests in this manner will allow a positive rather than a negative *Z*-score for the test statistic to be interpreted as supportive of this purpose.

The efficacy analysis will use a modified Haybittle-Peto stopping rule based on the stratified log-rank *Z*-score. If that *Z*-score equals or exceeds 3.2905 in absolute value, for either two-arm comparison, the difference between treatment arms would be declared statistically significant at the two-tailed 0.001 level of significance. In that case it may be reasonable for the DMC to initiate a conversation about stopping the trial on ethical grounds. Assuming this is not the case and the trial continues to its planned end, the *Z*-score criterion for declaring significance at the 5% level at the end of the trial will be increased in magnitude to plus or minus 1.962 in order to preserve the overall type I error rate for the trial at 0.05.

For the futility analysis, the DMC will be provided with the numerical value of the stratified log-rank *Z*-score. The futility analysis uses a one-sided futility stopping criterion of $Z = -0.50$. That is, if the *Z*-score is less than or equal to −0.50, an investigation will be initiated to consider stopping the trial for futility or discontinuing one of the single-agent treatment arms. The futility stopping criterion of $Z = -0.50$ is consistent with a conditional power of less than 20%. That is, assuming between 44 and 60 FAP-related events have occurred by trial end in either of the two-arm comparisons (where between 52 and 55 are expected), if the log-rank critical ratio *Z*-score were equal to −0.5 (or less) when one-half the expected total number of events had been observed (namely, 45 across all three arms), then under the design alternative hazard ratio of 2.3569, there would be no more than a 20% chance of declaring a significant benefit of the combination therapy compared to the single agent therapy if the trial were to continue to the planned end. In that case, it would be reasonable for the DMC to consider stopping or altering the trial on grounds of futility. The DMC will also be provided with the conditional power of the observed *Z*-score for each two-arm comparison.

The details of the efficacy and futility analysis are provided in Table 5.1 below which presents a schematic diagram of the procedure. Also Table 5.2 provides more precise conditional power values corresponding to the futility criterion $Z = -0.50$ as a function of the total number of events in either two-arm comparison. See Section 7.4 below for further details of the conditional power calculation.

Any numerical values generated from the futility analysis (such as *Z*-score, conditional power, etc.) must be treated as confidential by the DMC and Independent Statistician at the CRO. If the DMC recommendation is to continue the study as planned, such numerical values will not be forwarded or conveyed in any manner to the Steering Committee, sponsor, or any other parties.

Table 5.1 Interim Futility Analysis Details

Estimated Look Time Point	Description
When 45 primary endpoints have occurred, corresponding to when 50% of maximum trial information has been amassed	Efficacy criterion $Z = 1.962$ at terminal analysis. Futility criterion of $Z = -0.50$ or less at interim analysis. Type I error for exiting the upper efficacy boundary at or before terminal analysis = 0.024952.* Type I error for exiting the lower efficacy boundary at or before terminal analysis = 0.001928.* Total Type I error for end of study comparison = 0.026936.* Assuming 52 events at trial end in either two-arm comparison, power = 0.8705. Assuming 55 events at trial end in either two-arm comparison, power = 0.8881.

*Assume one interim analysis at ½ information time with upper efficacy boundary at $B_e = 3.2905$, lower efficacy boundary at $-B_e$, and one-sided futility boundary $B_f = -0.50$. Assume upper efficacy terminal criterion $C = 1.9602$ and lower efficacy terminal criterion $-C$. Then the probability of exiting the upper efficacy boundary at either interim or terminal analysis under the null hypothesis is given by

$$\Phi(-B_e) + \int_{B_f}^{B_e} \Phi(z - C\sqrt{2})\varphi(z)dz$$

and the probability of exiting the lower efficacy boundary at either interim or terminal analysis under the null hypothesis is given by

$$\Phi(-B_e) + \int_{B_f}^{B_e} \Phi(-z - C\sqrt{2})\varphi(z)dz,$$

where $\varphi(z)$ is the standard normal probability density function and $\Phi(z)$ is the standard normal cumulative distribution function.

Figure 5.1 Schematic Diagram for Efficacy, Futility, and Terminal Analyses

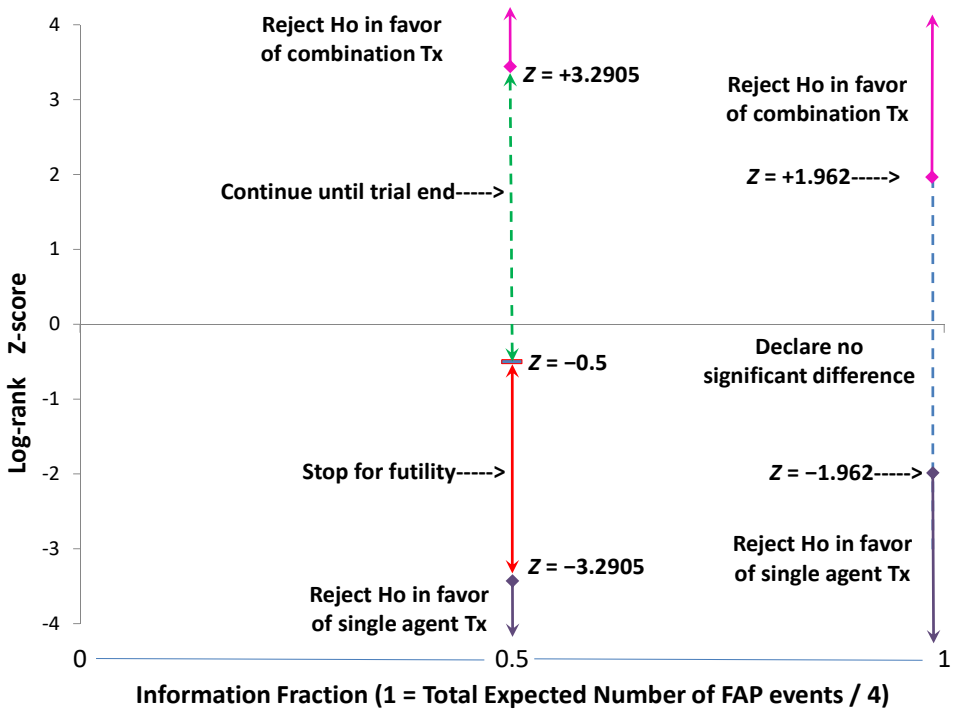

Table 5.2 Conditional Power at the Futility Boundary as a Function of Number of FAP Events at the Time of Futility Analysis, Using Futility Criterion $Z = -0.5$

Total FAP events in either two-arm comparison at time of futility analysis	Assumed number of FAP-related events in either two-arm comparison at trial conclusion	Conditional power
30	60	0.177
29	58	0.167
28	56	0.157
27	54	0.148
26	52	0.138
25	50	0.129
24	48	0.120
23	46	0.111
22	44	0.103

5.5 Statistical Methods

5.5.1 Demographic and Baseline Characteristics

Patients in the three populations (see Section 4.) will be summarized for demographic and baseline characteristics in a descriptive fashion. Namely, categorical and continuous-valued data will be displayed using standard summary statistics (e.g., frequency tables, n, means, medians,

standard deviations, and ranges). Data will be presented per group and overall.

Demographic features summarized will include age, gender, race, institution at which each patient registered, and country, among other features. Baseline characteristics will include laboratory values and disease-related characteristics, as well as any other relevant values. Categorical data will be compared among groups using chi-squared methods, while continuous-valued data will be compared using standard nonparametric methods (e.g., the Kruskal-Wallis test)[5]. Significance will be defined at the 0.05 level, unless otherwise noted. Thus p-values less than or equal to 0.05 will be declared significant.

5.5.2 Patient Disposition and Treatment Summaries

Subjects will be assigned for analysis to the treatment group to which they were randomized, regardless of whether the patients received any treatment.

Patient disposition and treatment will be summarized for the ITT and safety populations defined previously (see Section 4.2). Patient disposition will be consistent with the CONSORT criteria[6], and will include per treatment group enumeration of all patients randomized, the number deemed ineligible, the number of FAP-related events, and the number of study drop outs. These will be further described in subgroups such as drop outs due to adverse events, serious adverse events, administrative withdrawals for non-compliance for more than 90 days from randomization to month 36 or more than 105 days from randomization to month 42 or more than 120 days from randomization to month 48, withdrawals of consent for continued follow-up, withdrawals for other reasons, and the number lost to follow-up. Additional summaries will include reasons for patients discontinuing treatment and/or modifying treatment dosages. A listing of screened and ineligible patients along with the reason for each also will be summarized.

5.5.3 Analytic Methods for Time-to-Event Data

The analytic method for the primary analysis will be a time-to-event analysis using the stratified log-rank test, as previously described. The stratified Cox proportional hazards regression models will be used for secondary assessments[7]. Graphical analyses (log-minus-log plots) will be used to check the assumption of constant hazard ratios. See Section 7.3 below for further details of the calculation of the stratified log-rank statistic.

For the primary analysis, two stratified log-rank tests will be performed with treatment coded as a binary value (i.e., 0 or 1). Time to event curves will be displayed using the method of Kaplan and Meier[8]. Additional analyses involving the overall 3-treatment group comparison, and use of additional study populations (see Section 4.) for the two pairwise treatment comparisons, will be performed as supplemental analyses.

If an FAP-related event occurs, that patient will be said to have an observed or uncensored event and will be considered a treatment failure. If a subject withdraws, that subject will be treated as a censored observation as of the last recorded clinic visit (endoscopic disease assessment).

If a subject has not progressed or is not known to have died at the date of analysis cut-off, time to first FAP-related event will be censored at the date of the last adequate endoscopy procedures before the cut-off date. Similarly, if a subject discontinues study participation due to toxicity and begins receiving other therapy, the time to FAP event will be censored at the date of the last adequate endoscopy procedure.

If a subject has two or more missing assessments, time to first FAP-related event for the subject will be censored at the time of last adequate evaluation prior to the missing assessment.

If a subject has no baseline assessment, time to first FAP-related event for the subject will be censored at the date of randomization.

See Section 5.5.6 below for further details of handling missing data. Every effort will be made to minimize the occurrence of censoring and missing data.

Prior to the primary analysis, balance will be assessed between the three arms in terms of key potential confounders measured at the baseline visit. If any of these variables is found significantly out of balance across the three groups using a 2 degree of freedom test of homogeneity at the 0.01 level of significance, it will be incorporated into a sensitivity analysis using a stratified Cox model including that term in addition to the treatment arm. The primary result for the trial will be the unadjusted stratified log-rank test. The covariate-adjusted score test (adjusted stratified log-rank test) will serve only as a secondary analysis to aid in the interpretation of the primary result.

5.5.4 Analytic Methods for Categorical or Continuous-Valued Secondary Outcome and Safety Data

For categorical data, comparisons will be made between treatment groups using standard chi-squared techniques as the primary approach. In particular, the Cochran-Mantel-Haenszel one degree of freedom test will be used to reflect the stratified randomization. Exact p-values and 95% confidence intervals by the point-probability method will be reported[1].

For continuous endpoints, standard analysis of covariance (ANCOVA) methods will be used as the primary approach to compare treatment groups at end of treatment with the following covariates: baseline value, binary indicator variables for the two highest-risk stratification levels used in the randomization (using the lowest-risk, i.e., rectum/pouch polyposis, group as the reference stratum), and a binary treatment indicator (1=combination treatment, 0=single treatment).

For ordered categorical data, a Kruskal-Wallis nonparametric test for ordered categorical response will be used to compare treatment groups[5].

5.5.5 Assessment of Toxicities

Treatment-emergent adverse events will be enumerated and analyzed according to the incidence, intensity, type of adverse events, and clinically significant changes in the patient's physical examination findings, vital signs and clinical laboratory results. Safety variables will be tabulated and presented for all patients in the safety and per-protocol populations as defined previously.

Adverse events will be graded and coded using the NCI Common Terminology Criteria for Adverse Events (CTCAE, Version 4.03). Treatment-emergent events will be tabulated, where treatment-emergent is defined as any adverse event that occurs after administration of the first dose of study drug and through 30 days after the last dose of study drug, or any event that is present at baseline and continues after the first dose of study treatment but worsens in intensity. Events that are considered related to treatment (possibly, probably or definitely drug-related) will also be tabulated separately. Tables that enumerate adverse events by severity will be provided. Deaths, serious adverse events and events resulting in study discontinuation will be tabulated in data listings including additional relevant information on each patient. Tables will be presented both overall (all arms combined), by each treatment group separately, and by cell. Where appropriate, statistical comparisons between treatment arms will be provided using the above-mentioned methods for analysis of categorical data.

5.5.6 General Procedures for Handling Missing Data

Every reasonable effort will be made to continue follow-up of all study participants, including those who discontinue randomized therapy, to prevent data loss. It is recognized that missing values represent a potential source of bias in a clinical trial and so every effort will be undertaken to fulfill all the requirements of the protocol concerning the collection and management of data.

For the primary time to event analysis, the only possible patient outcome is an observed FAP-related event or a censored observation. Participants who are lost to follow-up will be censored at the time their status is last known, based upon data collected at the last recorded clinic visit. For patients who may have missed a study visit, every effort will be made to obtain endoscopic results at their close-out visit and those endoscopy results will be used for the primary analysis

Secondary analysis data include the presence of a specific genetic mutation, and urinary metabolite concentrations (see Section 3.2.3). The main analysis of the secondary objectives will include collected data only, without imputing or weighting data to compensate for missing data. For sensitivity analyses involving secondary endpoints with missing data, we will use the last observation carried forward (LOCF) method to complete the missing data. Any sensitivity analysis that incorporates LOCF will be clearly noted. Sensitivity analyses of these data will be performed to explore study results more fully, in a manner consistent with ICH Guidance "*E9 Statistical Principles for Clinical Trials (February, 1998)*".

All available efficacy and safety data will be included in data listings and tabulations. Data that are potentially spurious or erroneous or appear as outliers will be examined using standard data management operating procedures, prior to database lock and statistical analysis. These procedures will be fully described in the study report.

5.5.7 Subgroup Analyses

Subgroups will be analyzed in the spirit of exploratory analyses including but not limited to the various study populations (see Section 4) and separately within each disease-prognosis stratum.

5.6 Health Related Quality of Life (HRQoL) Statistical Methods

For this study four (4) instruments to measure HRQoL and patient preferences or utilities will be administered to subjects at baseline and at 3, 6, 12, 18, 24, 30, 36, 42 and 48 months post-enrollment/end of treatment. These instruments include the EORTC QLQ-C30, EORTC QLQ-CR29, EQ-5D, and a modified Cancer Worry Scale.

- The EORTC QLQ-C30 is a self-administered quality of life questionnaire[9] with multi-dimensional scales. It consists of both multi-item scales and single item measures, including five functioning domains, a global quality of life domain, three symptom domains and six single items.
- The EORTC QLQ-CR29 gastrointestinal/colorectal sub-module[10] is composed of 4 functional and 18 symptom related sub-scales. The 4 functional scales include body image, weight, anxiety and sexual function. The symptom related scales include single item and multi-item questions concerning stool frequency, bleeding and mucous discharge, stool leakage, abdominal bloating, flatulence, embarrassment and site-specific pain among others.
- The EuroQoL EQ-5D is a standardized instrument for use as a measure of health outcome and is applicable to a wide range of health conditions and treatments[11, 12]. It provides a simple descriptive profile and a single index value for health status.

- The Cancer Worry Scale[13] is a brief psychometric instrument that was designed to assess both the frequency of worrying about "getting cancer some day" and measuring the impact of worry on mood and performing daily activities. This scale was originally developed by Caryn Lerman and her colleagues to study breast cancer and has been modified for use in this FAP trial.

The validity and reliability of both the QLQ-C30 and the QLQ-CR29 questionnaires have been studied by the EORTC Study Group on Quality of Life and both instruments will be scored according to the EORTC Scoring Manual and analyzed accordingly. For each single item or multi-item sub-scale, a linear transformation will be applied to standardize raw scores to range between 0 and 100. HRQoL secondary endpoints will include all single item or multi-item sub-scales from both the EORTC QLQ-C30 and QLQ-CR29 and patients will be considered as deteriorated (or improved) for a given single item or multi-item sub-scale if their change score from baseline was 10 points or more on the standardized scale.

Patient preferences (or utilities) will also be assessed using the EuroQoL EQ-5D. Preference weights among the treatment arms will be determined using the EuroQol EQ-5D assessment of individual health states[11, 12]. Quality-adjusted survival among the three treatment arms will be generated by multiplying the utility value by the amount of time spent in a specified health state.

The modified version of the Cancer Worry Scale will also be administered and it will be scored according to the guidance provided by Lerman *et al*[13].

HRQoL data will be obtained while patients are receiving treatment. At the time of an FAP-related event (primary outcome), additional long-term clinical follow-up and QoL data will not be obtained as part of this trial. Hence, HRQoL trends comparing the nine subsets will be obtained, but comparative longitudinal analyses defining the impact of an FAP-related event on QoL will not be feasible until subsequent long-term studies are performed.

5.7 Dietary Assessment

The Food Frequency Questionnaire (FFQ) is the most common dietary assessment tool used in large epidemiologic studies of diet and health. The self-administered FFQ booklet asks participants to report the frequency of consumption and portion size of approximately 125 line items over a defined period of time (e.g. the last month; the last three months). Each line item is defined by a series of foods or beverages. Additional questions on food purchasing and preparation methods enable the analysis software to further refine nutrient calculations. The FFQ was developed by the Nutrition Assessment Shared Resource (NASR) of the Fred Hutchinson Cancer Research Center. NASR periodically updates its standard FFQ to reflect U.S. food consumption patterns and major changes in the market place[14, 15]. Data from the FFQ will be analyzed using a polyamine database[16-18] and will calculate the average daily levels of putrescine, spermidine, and spermine in the diet. Dietary assessments via the FFQ will be obtained at baseline, months 12, 24, 36, 42 (only if end of treatment), and 48/end of treatment for subjects at North American (U.S. and Canada) sites only. Clinical study sites outside of North America will not be included because the foods on the North American food frequency questionnaire are not the same as those widely consumed in Europe and elsewhere.

6. RANDOMIZATION ALGORITHM AND STRATIFICATION FACTORS

At least 150 eligible patients will be randomized in this study. Patients will be randomized to one of three treatment groups in equal proportions (i.e., 1:1:1 randomization): 1) CPP-1X plus sulindac, 2) CPP-1X-placebo plus sulindac, 3) CPP-1X plus sulindac placebo.

A stratified randomization procedure will be used with stratification based on FAP-related time-to-first-event prognosis. The event prognosis groups are represented by 1) best (i.e., longest projected time to first FAP-related event) - rectal/pouch polyposis, 2) intermediate - duodenal polyposis, and 3) worst - pre-colectomy. If a subject has two or more of these disease sites, the most severe prognosis stratum will be assigned for randomization (e.g. worst > intermediate > best). Since an individual may have more than one disease site involved, the trial will assess time to any defined FAP-related event in the patient as a whole. In order to minimize potential treatment arm imbalance a centralized randomization process will be used to balance among treatment groups within prognostic strata.

7. OTHER ISSUES AND FURTHER DETAILS

7.1 Statistical Software Used in Data Analysis

All analyses will be performed using SAS statistical analysis software version 9.1 or later. If other software is used (i.e. WinNonPop for population pharmacokinetics, SAS macros for futility analysis, etc.), it will be clearly described in the clinical and statistical study reports.

7.2 Draft Tables, Listings and Figures

The TLF (Tables, Listings, Figures) templates and shells are noted as Ver. 2.0, October 6, 2014 (SAP Ver. 3.0 September 30, 2014) based on revision provided from Data Management/Statistical Group at Ockham now Chiltern. These draft templates have been removed as an Appendix to the amended SAP and are available upon request.

7.3 Details of Calculating the Stratified Log-rank Z-Score

Patient follow-up will be analyzed in continuous time, although it is recognized that FAP event detection will cluster around scheduled study visits, at months 6, 12, 18, 24, 30, 36, 42 and 48 and event times may be tied as follow-up time is measured only to the nearest day. Censorings are possible at the baseline visit (month 0) for patients who dropped out of the trial before their first follow-up visit. The primary test statistic will be the stratified log-rank statistic. Each two-arm comparison will be performed separately. This will be implemented in SAS by specifying the TIES=DISCRETE option in PROC PHREG. The single agent treatments, respectively, will be coded as 1 and the combination treatment will be coded as 0. The square root of the score test chi-squared statistic will be calculated and the *sign* of the estimated log hazard ratio (+ if the HR>1 and − if the HR<1) will be attached. This results in the stratified log-rank *Z*-score. In symbols, $Z_k = sign(\hat{\beta}) \times \sqrt{X^2}$, where $\hat{\beta}$ denotes the estimated log hazard ratio coefficient for treatment *A* in the Cox model and where X^2 denotes the SCORE TEST chi-squared statistic reported within the "Testing Global Null Hypothesis: BETA=0" table at the top of the PHREG results section.

The following sample PHREG procedure could be used in SAS to derive the stratified log-rank *Z*-score:

```
proc phreg data=TwoArmData;
model time*censor(1) =  Treatment / rl ties=discrete;
```

```
strata Stratification_Factor;
title "Stratified Cox Regression Analysis for Single Agent vs Combination";
run;
quit;
```

NB: The model includes the variables time, censor, and treatment, only.

7.4 Details of Calculating Conditional Power

Conditional power is given by the following formula[20].

$$\text{Conditional power} = \Phi\left(\frac{Z_k\sqrt{I_k} - z_{1-\alpha/2}\sqrt{I_K} + \theta(I_K - I_k)}{\sqrt{I_K - I_k}}\right),$$

where

Z_k is the value of the stratified log-rank test statistic from the observed data at interim analysis;
$z_{1-\alpha/2}$ is the critical value for a two-tailed test at α=0.05, namely +1.962 adjusted for the interim efficacy analysis;
I_K = the information number expected at trial end = ¼ times the total expected number of events;
I_k = the information number at futility analysis = ¼ times the observed number of events at interim analysis;
θ is the log hazard ratio at the design alternative, log 2.3569 = −log 0.42428 = +0.85736; and
$\Phi(\cdot)$ is the standard normal cumulative distribution function.

Example. Suppose at the time of interim analysis, the number of FAP-related events are as follows: For single-agent treatment *A*, 20; for single-agent treatment *B*, 16; and for combination treatment *C*, 9. Then in the two-arm comparison of *A* versus *C*, there would be a total of 29 events, so the information number is I_k=29/4=7.25. We assume that there will be double this number for the corresponding comparison at the end of the trial; thus I_K = 14.5. Suppose the *Z*-score at the futility analysis happens to just equal the futility criterion, i.e., suppose Z_k = −0.50. Then applying the above formula yields conditional power of

$$\Phi\left(\frac{-0.5\sqrt{7.25} - 1.962\sqrt{14.5} + 0.85736(14.5 - 7.25)}{\sqrt{14.5 - 7.25}}\right) = 0.1670\ .$$

This is the conditional power *at the futility boundary*. Suppose, however, that the observed *Z*-score at the futility analysis were actually +0.50 instead of −0.5. Then the −0.5 in the first term of the above expression would be replaced by Z_k=+0.50, and the resulting *observed* conditional power would be 0.5135, a non-futile result.

Note that the probability of a significant *negative* result at the terminal analysis, given that there is no stopping for futility at the interim analysis, is not included in the above expression. That is because practical interest for futility analysis resides only in the conditional probability of a significant benefit of the combination treatment. In any case, the term which has been omitted is numerically negligible.

8. REFERENCES

1. Fleiss, J., B. Levin, and M.C. Paik, *Statistical Methods for Rates and Proportions*. 3rd Edition ed. 2003: New York: John Wiley & Sons. See generally Section 2.7 and especially Section 10.3 at page 243.
2. Hochberg, Y.A., *A Sharper Bonferroni procedure for multiple tests of significance.* Biometrika, 1988. **75**: p. 800-802.
3. Freedman, L.S., *Tables of the number of patients required in clinical trials using the logrank test.* Stat Med, 1982. **1**(2): p. 121-9.
4. Schoenfeld, D., *The asymptotic properties of nonparametric tests for comparing survival distributions.* Biometrika, 1981. **68**(1): p. 316-319.
5. Lehmann, E.L., *Nonparametrics: statistical methods based on ranks.* 1975, San Francisco: Holden-Day.
6. Schulz, K.F., et al., *CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials.* PLoS Med, 2010. **7**(3): p. e1000251.
7. Cox, D.R., *Regression models and life tables.* J Royal Stat Soc, 1972. **Series B**: p. 34182-34202.
8. Kaplan, E. and P. Meier, *Nonparametric estimation from incomplete observations.* J. Amer. Stat. Assoc., 1958. **53**: p. 457-481.
9. Aaronson, N.K., et al., *The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.* J Natl Cancer Inst, 1993. **85**(5): p. 365-76.
10. Whistance, R.N., et al., *Clinical and psychometric validation of the EORTC QLQ-CR29 questionnaire module to assess health-related quality of life in patients with colorectal cancer.* Eur J Cancer, 2009. **45**(17): p. 3017-26.
11. Brooks, R., *EuroQol: the current state of play.* Health Policy, 1996. **37**(1): p. 53-72.
12. Drummond, M.F., et al., *Methods for the Economic Evaluation of Health Care Programme*. 2nd ed. ed. 1997, New York: Oxford University Press.
13. Lerman, C., et al., *Psychological side effects of breast cancer screening.* Health Psychol, 1991. **10**(4): p. 259-67.
14. Kristal, A.R., A.L. Shattuck, and A.E. Williams. *Food frequency questionnaires for diet intervention research*. in *17th National Nutrient Databank Conference*. 1992. Baltimore, MD.
15. Neuhouser, M.L., et al., *Validity of short food frequency questionnaires used in cancer chemoprevention trials: results from the Prostate Cancer Prevention Trial.* Cancer Epidemiol Biomarkers Prev, 1999. **8**(8): p. 721-5.
16. Zoumas-Morse, C., et al., *Development of a polyamine database for assessing dietary intake.* J Am Diet Assoc, 2007. **107**(6): p. 1024-7.
17. Raj, K.P., et al., *Role of dietary polyamines in a phase III clinical trial of difluoromethylornithine (DFMO) and sulindac for prevention of sporadic colorectal adenomas.* Br J Cancer, 2013. **108**(3): p. 512-8.
18. Zell, J.A., et al., *Associations of a polymorphism in the ornithine decarboxylase gene with colorectal cancer survival.* Clin Cancer Res, 2009. **15**(19): p. 6208-16.