Statistical Analysis Plan Amendment 1
29 June 2015

BNIT-PRV-301

A Randomized, Double-blind, Phase 3 Efficacy Trial of PROSTVAC VF ± GM-CSF in Men With Asymptomatic or Minimally Symptomatic Metastatic, Castrate-Resistant Prostate Cancer

Bavarian Nordic, Inc.

TABLE OF CONTENTS

List of Abbreviations

ALP	Alkaline Phosphatase
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
ATC	Anatomical-Therapeutic-Chemical
AWE	Alive Without Event
BNI	Bavarian Nordic, Inc. Formerly known as BNIT
BPI-SF	Brief Pain Inventory – Short Form
BUN	Blood urea nitrogen
CK	Creatine Kinase
CK-MB	Creatine Kinase - Myocardial Band
CRP	C-Reactive Protein
CTC	Circulating Tumor Cell
CTCAE	Common Terminology Criteria for Adverse Events
CTMS	Clinical Trial Management System
DMC	Data Monitoring Committee
ECG	Electrocardiogram
eCRF	Electronic case report form
EOT	End-of-treatment
FACT-P	Functional Assessment of Cancer Therapy - Prostate
FAS	Full Analysis Set
GM-CSF	Granulocyte Macrophage Colony-Stimulating Factor
HBSAG	Hepatitis B Surface Antigen
HCT	Hematocrit
HCV	Hepatitis C Viral Load
HGB	Hemoglobin
IAP	Interim Analysis Plan
INR	International Normalized Ratio
ITT	Intent-to-Treat
LDH	Lactate dehydrogenase
LTFU	Long-term Follow-up

mCRPC	Metastatic castrate-resistant prostate cancer
MedDRA	Medical Dictionary for Regulatory Activities
NCI	National Cancer Institute
OS	Overall Survival
PAP	Prostatic Acid Phosphatase
PK	Pharmacokinetic
PPACT	Post-progression Anticancer Therapies
PSA	Prostate Specific Antigen
PT	Preferred Term
PTT	Partial Thromboplastin Time
QOL	Quality-of-life
RBC	Red Blood Cells
SAE	Serious Adverse Event
SC	Subcutaneous
SIP	Statistical Inference Plan
SOC	System Organ Class
TEAE	Treatment Emergent Adverse Event
WBC	White Blood Cells
WHO	World Health Organization

Revision History

SAP Version	SAP Date	Protocol Version	Reason for Change
Final	28 February 2013	Amendment 5	Initial version
Amendment 1	29 June 2015	Amendment 6	Updated per BNIT-PRV-301 protocol amendment 6

1. Introduction

PROSTVAC™ (PROSTVAC V and PROSTVAC F; PROSTVAC VF) is a novel candidate prostate cancer immunotherapy for the treatment of prostate cancer. It is a viral vector based product that is administered in seven subcutaneous vaccinations over a five month period. In a randomized controlled Phase 2 trial, PROSTVAC therapy was associated with a prolongation of survival in men with metastatic castration-resistant prostate cancer.

PROSTVAC is a PSA (prostate-specific antigen)-based immunization strategy. It is intended to generate immune responses to prostate specific antigens and prostate cancer cells. It uses poxviral vectors to introduce modified PSA to the subject in an immunogenic manner to break self-tolerance, and thereby induce immune responses directed against prostate cancer cells.

PROSTVAC is comprised of two component viral vectors; a recombinant vaccinia (PROSTVAC V) and a recombinant fowlpox (PROSTVAC F) virus to be used sequentially in a heterologous prime-boost vaccination regimen. Both viruses contain four human genes: prostate-specific antigen (PSA) and three genes encoding human immunological costimulatory molecules: B7.1, intercellular adhesion molecule-1 (ICAM-1), and leukocyte function-associated antigen-3 (LFA-3) (designated TRIad of COstimulatory Molecules or TRICOM). The simultaneous expression of PSA and TRICOM enhances the immune response. The PSA transgene is modified at one amino acid (I155L), to enhance binding and immunogenicity of a particular peptide epitope in HLA-A2-expressing subjects.

2. Objectives

This is a randomized, double-blinded, placebo-controlled, multi-center, three-arm study. Subjects will be randomized with equal probability to one of three arms, designated as P, V, and VG for statistical purposes. A general description of the intended interventions to be applied to the subjects randomized to these three arms is: Placebo PROSTVAC + Placebo Granulocyte Macrophage Colony-Stimulating Factor (GM-CSF) for Arm P, PROSTVAC + Placebo GM-CSF for Arm V, and PROSTVAC + GM-CSF for Arm VG.

2.1. Primary Objective

The primary efficacy objective is to ascertain whether there is evidence for either Arm V or Arm VG to have superior survival outcome relative to Arm P.

2.2. Secondary Objective

To ascertain whether a greater proportion of subjects randomized to Arm VG or to Arm V remain event free at six months (or early termination) as compared to the subjects randomized to Arm P.

2.3. Exploratory Objectives

The Exploratory Objectives are:

- To evaluate whether HLA-A2 positive subjects have a larger survival effect size as compared to HLA-A2 negative subjects.
- To model the radiological disease status over the first year of study participation.
- To assess the role of post-treatment anti-cancer therapies as an alternative explanations for observed survival differences.
- To compare the arms with respect to immune responses to immunizing antigen (PSA), as well as to non-vaccine containing prostate antigens, and tumor-associated antigens, and to assess whether immune responses are prognostic.
- To compare the arms with respect to circulating tumor cell (CTC) levels.
- To model the role of arm, potential baseline prognostic factors, including baseline subject attributes and circulating tumor cells, immune responses, and the use of post-progression anti-cancer therapies as explanatory factors of survival.
- To model the relationship between tumor regression and growth rates and duration of survival.

3. Investigational Plan

3.1. Overall Study Design and Plan

BNIT-PRV-301 is a randomized, placebo-controlled, multi-center, Phase 3 efficacy trial of PROSTVAC in men with asymptomatic or minimally symptomatic, metastatic, castration-resistant prostate cancer. The study will consist of three periods: a Screening phase, followed by Treatment, and subsequently Long-term Follow-up (LTFU) phases.

Following the completion of the Treatment phase (after the last dose of PROSTVAC-V/F +GM-CSF or placebo or early termination of treatment) of the trial, all subjects will automatically enter the LTFU phase with study visits occurring every six months. During LTFU subjects will receive standard-of-care treatment as determined by the Principal Investigator. The appropriate timing and the nature of any subsequent treatment will be at the discretion of the Principal Investigator. All treatments for mCRPC subsequent to the Treatment phase of the trial will be documented in the electronic case report form (eCRF). This comprehensive data will be collected until study closure or death.

A detailed schedule of assessments can be found in the protocol.

3.2. Study Endpoints

3.2.1. Primary Endpoint

The primary endpoint is overall survival (OS). OS is defined as the time between the date of randomization and the date of death due to any cause. Subjects who do not experience death or the competing events of "definite" loss to follow-up or withdrawal of consent will be right censored at the date of last contact.

OS = Date of death/competing event/censoring - date of randomization+ 1.

3.2.2. Secondary Endpoint

The secondary efficacy endpoint for the study is the proportion of event-free subjects (AWE) (radiographic progression, pain progression, chemotherapy initiation, or death) at six months (or early termination) compared to placebo.

Progression events will be defined as:

- Radiographic progression (per central radiology assessment)
    - Two new lesions on bone scan, new metastases on CT scans*, or an increase size of nodal lesions per RECIST 1.1.
- Pain progression**
    - Introduction of scheduled opioid narcotics for cancer-related pain control. Scheduled opioid narcotics are defined as continuous daily use lasting more than 14 days.
- Initiation of chemotherapy for prostate cancer
- Death

* For subjects with allergies to contrast agents, MRIs for abdomen and pelvis may be performed. Only in cases where MRI is unavailable may a non-contrast CT be performed.

** The start date for an event of pain progression is defined as day 1 of first opioid regimen lasting more than 14 days.

Radiographic progression will be assessed by bone scan or CT at Screening, Week 13, Week 25 (EOT), and early termination scans. For secondary endpoint analysis, Screening scans and Week 13, Week 25 (EOT), and early termination scans will be compared by central radiology review. Bone scan and CT data will be archived up to five years past the end of the trial.

The secondary efficacy endpoint is a binary assessment for AWE at six months. Subjects without an event prior to six months will have an evaluation for event at six months. Subjects without event who are not evaluated at six months will be assumed to have had an event. Thus, each subject will be defined as AWE = 1 if the subject is alive and free of event at six months and AWE = 0 otherwise. AWE will have a realization for every subject and therefore is amendable to an intent-to-treat analysis.

The analysis of AWE is less subject to ascertainment anomalies compared to the analysis of time-to-progression because the six month assessment brings all subjects to the same level of ascertainment.

3.2.3. Exploratory Endpoint

The exploratory efficacy endpoint for the study is the proportion of subjects Alive Without Event (AWE) (radiographic progression, pain progression, chemotherapy initiation, or death) at the 6 month LTFU visit compared to placebo. For all subjects alive at the Week 25 (EOT) visit we will assess AWE at the 6 month LTFU visit. Subjects with pain progression or initiating

chemotherapy for prostate cancer at the Week 25 (EOT) visit will be counted as having an event at the 6 month LTFU visit.

Progression events at the 6 month LTFU visit will be defined as:

- Radiographic progression (per central radiology assessment) taking the Week 25 (EOT) visit scan as baseline
    - Two new lesions on bone scan, new metastases on CT scans*, or an increase size of nodal lesions per RECIST 1.1.
- Pain progression**
    - Introduction of scheduled opioid narcotics for cancer-related pain control. Scheduled opioid narcotics are defined as continuous daily use lasting more than 14 days.
- Initiation of chemotherapy for prostate cancer
- Death

* For subjects with allergies to contrast agents, MRIs for abdomen and pelvis may be performed. Only in cases where MRI is unavailable may a non-contrast CT be performed.
** The start date for an event of pain progression is defined as day 1 of first opioid regimen lasting more than 14 days.

This will attempt to define effects on secondary progression, and characterize delayed effects of investigational vaccine therapy.

3.3. Study Treatment

The trial interventions will consist of a single subcutaneous (sc) immunization of PROSTVAC V or placebo in Week 1, followed by six PROSTVAC F or placebo immunizations administered in Weeks 3, 5, 9, 13, 17, and 21. Each immunization will be accompanied by administration of sc GM-CSF or placebo on the day of immunization and for the subsequent three days (sc injection within 5 mm of the original PROSTVAC VF or placebo injection site). The Treatment phase of this trial lasts for five months and is followed by an End-of-Treatment (Week 25)/Early Termination visit. During the Treatment phase, subjects will continue to be treated with vaccination therapy through PSA and/or radiological progression.

3.4. Dose Adjustment

See the Dose Reductions section of the protocol.

4. General Statistical Considerations

Continuous data will be described using descriptive statistics as follows: number of observations (n), mean, median, standard deviation, minimum, and maximum. All minimum and maximum values will be displayed with the same number of decimal places relative to the raw data, the mean and median will be displayed with one additional decimal place, and the standard deviation will be displayed with two additional decimal places.

Data listings will be displayed by subject identification number. Summary tables will be displayed by treatment arm for all subjects. For those summary tables in which baseline and change from baseline measurements will be presented, if time is not available the last observed measurement on or before the date of initial dose of study product will be considered the baseline measurement. If date and time are available the last observed measurement on or before the date and time of initial dose of study product will be considered the baseline measurement.

Study Day 1 is defined as the date of first dose of study product. For visits (or events) that occur on or after the first dose of study product, study day is defined as visit date (event date) – date of first dose of study product + 1. For visits (or events) that occur before the first dose of study product, study day is defined as visit date (event date) – date of first dose of study product. There is no Study Day 0.

For listings that include the derivation of "days since last dose," this is defined as event date – date of last dose. Events that occur on the same day as the last dose of study product will therefore be described as occurring zero days from the last dose of study product.

When categorical data are presented, the percentage will be suppressed when the count is zero in order to draw attention to the non-zero counts. The denominator for all percentages, unless otherwise specified, will be the number of subjects in each treatment arm or overall.

Only nominal/scheduled visit data will be reported in the tables. All visits, including unscheduled visits, will be presented in the listings. Note that all scheduled visits should occur within ± 3 days of the protocol-specified visit schedule.

4.1. Sample Size

4.1.1. Arm P Reference Data

The Arm P reference data for planning this trial is specified as a median survival of 22 months based on the placebo arm of Dendreon D9902B. Using exponential assumptions the corresponding hazard rate is 0.3781 and the corresponding three-year survival is 32.2%.

4.1.2. Delay of Vaccine Effect

There is evidence, both from the PROSTVAC phase 2 study and other prostate cancer trials, as well as trials of therapeutic vaccines in other cancers, that there exists the possibility of a delay in the effect of the experimental vaccine. Taking into account the possibility of a delay in effect precludes using trial size computation methods that assume proportional hazards because the proportional hazards assumption is not valid if there is delayed effect. The existence of a delayed effect means that early deaths will not differentiate between the arms and therefore the number of deaths required for the same power will be greater. Thus, the use of methods that do not take into account delay of effect could lead to underestimating the power of the trial. Simulations were used for both delayed effect scenarios and scenarios where delayed effect was not assumed. **Appendix A** to this document describes the simulation methodology.

4.1.3. Effect Size for Survival

The hazard ratio in the specific alternative hypotheses used to compute trial size is 0.68, that is, a 32% smaller hazard rate in Arm VG or Arm V as compared to Arm P. This hazard ratio is based on hazard ratio estimates from the previously completed randomized phase 2 trial with an allowance that the phase 2 trial likely provides an over-optimistic estimate of the true hazard ratio. The hazard ratio estimate from the phase 2 trial was 0.56 when delayed effect is not taken into account, and was approximately 0.5 when delayed effect is taken into account (depending on the quantification of the delayed effect – see next paragraph).

Proportional hazards regression with time-dependent covariate implementation of delayed effect was used to obtain estimates of the ultimate hazard ratio following delayed effect in the previously completed randomized phase 2 study. It was found that estimates of this ultimate hazard ratio were reasonably uniform over a range of delayed effect specifications and that having delayed effect in the model resulted in a better fit to the survival data than models without delayed effect. The delayed effect estimate of the hazard ratio is approximately 0.5. The shape of the delay of effect used to plan this trial is based on that observed in the phase 2 study, but with an allowance that the phase 2 trial likely provides an over-optimistic estimate of the true hazard ratio.

The delay of effect used in the simulations done to plan this trial is based on that observed in the randomized phase 2 study. This delay of effect is characterized as follows: the hazard ratio is initially unity (no effect) for four months, and over the next four months transitions linearly to the hypothesized 0.68.

4.1.4. Power and Significance Level

The trial is planned so that a between-arm logrank test for all subjects will have the target power. The target power taking into account the possibility of a delay in effect is specified to be 85% or greater. The two comparisons will be done with Bonferroni correction to the overall type I error probability so that the type I error probability for each will not exceed one-sided 0.0125 (=0.025/2).

4.1.5. Number of Subjects and Required Number of Deaths

The number of subjects per arm is set to be 400 (Approximately 1200 total subjects). The accrual rate distribution is 19 months of linear ramp up followed by 11 months of constant rate of accrual (2.5 years to accrue approximately 1200 subjects). Note: The accrual distribution assumptions have been updated from the assumptions used originally to plan this trial and are based on current accrual experience. The table below has been revised to reflect these new accrual assumptions and there is no material modification to power, hypothesized hazard ratio, or any other important aspect of the trial design. As expected, the revised computations do change the projections for time to analysis. The assumptions used in **Table 1** match the assumptions used in the interim analysis planning computations outlined on pages 23 and 24. Please see **Appendix B** for more detail around the 19 month linear ramp followed by 11 months constant rate of accrual.

The trial is planned as an event driven trial. Simulation is used to compute the required number of deaths after setting other specifications. **Appendix A** to this document describes the simulation methodology. **Table 1** below shows the results of the simulations.

Table 1: Simulation Results

Case #	Delayed Effect	Specified # deaths	Time to specified deaths (y)	1-sided type I error probability estimate (%)	Power estimate (%)	Critical hazard ratio
1	No	340	3.4	1.17	90.2	0.784
2	Yes	340	3.3	1.13	44.5	0.784
3	Yes	534	5.1	1.23	86.1	0.823
4	No	534	5.2	1.12	98.8	0.823

Notes:

- Simulations are for the comparison of one of the experimental arms to control arm, that is, 800 subjects, 400 per arm, accrued over two years, with a 19 month ramp time followed by a constant accrual rate for 11 months (2.5 year total accrual).
- 10,000 replications for each case.
- Significance level is one-sided 0.0125. The column "One-sided type I error probability estimate (%)" from the simulations should be approximately 1.25%. Target power with delayed effect is 85% or more.
- The control arm data are generated under the exponential distribution with 22 month median survival.
- The experimental arm data under the specific alternative and when delayed effect (if applicable) are generated so that hazard rates are equal for 4 months (hazard ratio = 1), followed by a linear transition over next 4 months to target hazard ratio of 0.68, with hazard ratio = 0.68 from 8 months onward.
- The estimated type I error and power are the percent rejection of the null hypothesis when the data are simulated under the null or specific alternative hypotheses, respectively. The estimated type I error probabilities in the table are sufficiently close to 1.25% considering they are estimated from simulations.
- The "Critical hazard ratio" is the maximum hazard ratio that resulted in rejection of the null. This is an estimate of the threshold of significance. At the end of the trial larger hazard ratios will not result in rejection of the null hypothesis.

Case # 1 simulates a trial without delayed effect and a target power of 90% or greater and corresponds to conventional (proportional hazards) trial sizing computations, where the usual formula computed a requirement for 340 deaths. Case # 2 is based on the required number of deaths computed for Case # 1 (340) but with the data in the simulations generated under the specified delayed effect. The power for an analysis at 340 deaths when the specified delayed effect is applicable is not sufficient (44.5%). Note that 340 deaths are realized sooner than for Case # 2 than Case # 1 because the experimental arm will have a higher early hazard rate in Case # 2 than in Case # 1.

Case # 3 represents a solution to the required number of deaths for a target power of 85% or more under the delayed effect specified. The number of deaths required is found to be 534, with time to realization of 534 deaths projected to be 5.1 years. Case # 4 illustrates the estimated power of a trial with 534 deaths at the final analysis but without delayed effect. The high power (98.8) is the result of early events being differentiating.

Thus, these simulations show that a final analysis at 534 deaths provides protection against unacceptable loss of power if the projected delayed effect is applicable, and yet assures that the type I error probability is protected. The trial's power will be 85% or greater if the delayed effect does not exceed that specified in the simulations.

4.2. Randomization, Stratification, and Blinding

Block randomization will be used to randomly assign subjects in a 1:1:1 manner. Randomization will be stratified according to the following two factors, PSA (equal and above or below 50 ng/ml) and LDH (equal and above or below 200 U/L). These two stratification factors will be used to create a four-level categorical stratification variable for the following combinations of the stratification factors: (PSA < 50 & LDH < 200), (PSA < 50 & LDH ≥ 200), (PSA ≥ 50 & LDH < 200), and (PSA ≥ 50 & LDH ≥ 200). This stratification variable will be used to stratify analyses as specified subsequently.

Unblinding will be performed only in the event of a medical emergency in which the Investigator determines that knowledge of the subject's treatment assignment is necessary for effective treatment. If unblinding is required, the Investigator must contact the Medical Monitor or designee who will record who requested the unblinding, for what purpose, and for which subject, and will notify the DMC in a timely manner.

4.3. Analysis Sets

Three analysis sets are defined for this trial. The Intent-to-Treat set (ITT) includes all subjects who are randomized. The full analysis set (FAS) includes all subjects initiating study intervention. The Safety analysis set includes all subjects initiating study intervention. The FAS and safety analysis set are the same.

All efficacy analyses will be performed on the ITT set and the FAS. Subjects will be analyzed and summarized based on the arm into which they were randomized. Efficacy analysis performed on the ITT analysis set will be considered to be the primary indicator of efficacy. Analysis performed on the FAS will be considered to be supportive.

All safety analyses will be performed on the Safety set and subjects will be analyzed and summarized based on the treatment they actually received.

4.4. Changes in Planned Analysis

The analysis as described in the protocol has been expanded to include the analysis of tumor regression and growth rates. Interim analysis has been added.

5. Subject Disposition

Subject enrollment will be summarized for each site by treatment arm and overall for the FAS.

Subject disposition will be summarized by treatment arm and overall for the FAS. Percentages for this table will be based on the total number of subjects in each treatment arm and overall.

The reasons for study treatment discontinuation will be summarized by treatment arm and overall. The reasons for study treatment discontinuation are listed on the treatment completion eCRFs.

6. Demographics and Baseline Characteristics

6.1. Demographics

A summary of demographics and baseline characteristics will be presented by treatment arm and overall for the FAS. The demographic characteristics consist of age, age category, sex, ethnicity, race, and prior smallpox vaccination status. Baseline height and weight will also be presented using descriptive statistics.

For subjects who provide date of birth, age will be calculated as the integer part of (Date of Informed Consent – Date of Birth + 1)/365.25. Age will also be categorized using the categories 18-<25, 25-<45, 45-<65, and 65 or over, and will be presented using frequencies and percentages. The number and percentage of subjects' ethnicity (Hispanic or Latino, Not Hispanic or Latino, or Not Reported) as well as reported race categories will also be reported. Reported race categories include White, Black, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and Multiracial.

All demographic and baseline characteristic data will be listed for each subject.

6.2. Medical History

6.2.1. General Medical History

Medical history will be presented as a data listing.

6.2.2. Prostate Cancer History

Prostate cancer history and curative intent treatment for prostate cancer will be summarized by treatment arm and overall for the FAS.

Years since initial diagnosis, Gleason sum at initial diagnosis and days since recurrence after curative intent treatment will be summarized using descriptive statistics. TNM status at initial diagnosis will be summarized using frequency counts and percentages.

Years since initial diagnosis is defined as the integer part of (date of first dose – date of initial diagnosis + 1)/365.25.

Days since recurrence after curative intent treatment is defined as date of first dose – date of first increased PSA value since curative intent treatment + 1.

Incomplete dates will be imputed as follows:

- If day is missing, then the day will be set to the 15th of the month.
- If month and day are missing, then the month and day will be set to July 1st.

Curative intent treatment dates for select treatments are captured for all subjects. The number of subjects receiving these treatments will be summarized using frequency counts and percentages. "Other" curative intent treatment will be collected and the number of subjects receiving these treatments will be summarized by treatment class and generic trade name, or by treatment class and drug name if generic trade name is missing, using frequency counts and percentages.

PSA values for the 6 months prior to dosing for each subject will be presented in the data listings.

6.3. Eligibility

Eligibility deviations consist of any violations of the inclusion and exclusion criteria as detailed in the protocol. Eligibility deviations and other protocol deviations captured in the CTMS will be presented in separate listings.

7. Treatments and Medications

7.1. Prior and Concomitant Medications and Non-Drug Therapies

All medications recorded on the eCRF will be coded using the World Health Organization (WHO) Drug Dictionary. Prior and concomitant medications will be summarized for the FAS by Anatomical-Therapeutic-Chemical (ATC) Class Level 4 and WHO Drug Substance preferred term (PT). Subjects may have more than one medication per drug class and preferred term. At each level of summarization, a subject is counted once if he/she reported one or more medications at that level.

Prior medications are defined as medications with a stop date occurring before the date of first dose. Concomitant medications are defined as medications with start dates occurring on or after date of first dose and before the date of last dose + 28 days. Any medication with start and stop dates that bracket the date of first dose will be summarized as both prior and concomitant medications. Medications with start dates occurring more than 28 days after the date of last dose will be defined as Follow-up medications.

For the purpose of inclusion in prior and/or concomitant medication tables, incomplete medication start and stop dates will be imputed as detailed in **Appendix C** of this document. Based on imputed start and stop dates, medications that clearly stopped before date of first dose

will be included in the prior medications table, and medications that clearly started on or after date of first dose and before the date of last dose + 28 days will be included in the concomitant medications table. All other medications will be included in both the prior and concomitant medications tables.

Prior and concomitant medications will be presented in a listing. Follow-up medications will be presented in a separate listing.

7.2. Study Product Exposure

Study product exposure will be summarized by treatment arm and overall for the FAS.

Number of PROSTVAC or placebo PROSTVAC injections administered, number of GM-CSF or placebo GM-CSF injections administered and duration of treatment in days will be summarized using standard descriptive statistics. The number of subjects receiving each scheduled dose will also be summarized using frequency counts and percentages.

Duration of treatment will be calculated as date of last dose – date of first dose + 1.

Dosing details will be presented in listings for all subjects

7.3. Post-Vaccination Prostate Cancer Therapy

Post-vaccination prostate cancer therapy will be collected and the number of subjects receiving therapy will be summarized by treatment class and generic trade name, or by treatment class and drug name if generic trade name is missing, using frequency counts and percentages.

8. Efficacy Analysis

8.1. Analytic Standards

8.1.1. Data Cutoff Date

Associated with each analysis is a data cutoff date. The key analyses will use the data cutoff date to censor event dates, and this may result in an event date turning into a censored date. Additional sensitivity analyses without data cutoff censoring will also be done when there has been data cutoff censoring, especially when data cutoff censoring results in loss of events.

8.1.2. One-sided P Values

Statistical programs most often report 2-sided P values, but the criterion for the key analyses described herein are based on 1-sided P values associated with 1-sided null hypotheses. The computation of a 1-sided P value requires that the effect size estimate be available and that the favorable direction be known, that is, the direction indicated in the associated alternative hypothesis.

Assume that P_2 is the 2-sided P value reported from a statistical program. The one-sided P value, P_1, is to be computed as follows: If the effect estimate is consistent with favorable outcome then $P_1 = P_2/2$, otherwise $P_1 = 1 – P_2/2$.

8.1.3. Assessing Clinical Site Homogeneity

It is important to assess clinical site homogeneity with respect to the primary and secondary outcomes. Following is the method that will be used to obtain a descriptive assessment of clinical site homogeneity.

Assume each clinical site is identified by a unique code. It is often the case that one or more clinical sites will fail to have at least one subject assigned to each arm. All such clinical sites are to be pooled into a new pseudo clinical site with a new unique code. This new pseudo clinical site is to be used in the following analyses. A forest graph with sites or a grouping of sites will be created where the effect estimate will be displayed along with the 95% confidence interval. This graph is most effectively displayed if the estimates are sorted by estimate size.

8.2. Primary Analyses (Survival)

The survival outcome from subjects in each experimental arm, Arm V and Arm VG, will be compared separately to the survival outcome from subjects in the control arm, Arm P. If either of these comparisons passes the statistical criterion specified previously, then the trial will be regarded as having met the statistical criterion for success. This section defines the statistical method and criterion for each of these survival analyses.

The survival analysis objective is to ascertain whether the survival time distribution for subjects randomized to the experimental arm is consistent with longer survival as compared to subjects randomized to the control arm. All randomized subjects will be included in the primary survival analysis according to the randomized arm (intent-to-treat).

8.2.1. Imputation of Missing Efficacy Data

In general, there will be no imputation of missing data in efficacy analyses other than the primary endpoint (overall survival).

Imputation of the primary endpoint may depend on the date of the first occurrence of an event associated with the endpoint. When the event has yet to occur the analysis will nonetheless require an analytical date be available for the analysis. These are known as censored or truncated dates. A censored date is the date of the most recent assessment for the event for a subject for which the event of interest has yet to occur and for which additional time may lead to observation of the event of interest. A truncated date is the date of a competing event that precludes the observation of the event of interest. The distinction between designating an analytical date as censored or truncated will depend on the purpose of the analysis and will be designated for each analysis. For example, if death is the event of interest and a subject is still alive and being followed for the date of death then the date last known alive is a censored date, whereas the date a subject withdraws consent for continued follow-up of vital status (a competing event) is a truncated date. The nature of the competing event will dictate whether the

competing event is permanent. For example, the date of death for a subject who has withdrawn consent may later become known passively from governmental death records.

Dates associated with the primary endpoint with missing day of the month will be completed by imputing the missing day of the month as the 15th day of that month, provided another date to be used in the computation of an interval is not in the same month. If the day of the month is missing and the other date is in the same month, then the missing day of the month will be taken as half the distance to the beginning or end of the month (depending on expected sequence). If both dates to be used in the computation of an interval are missing day of the month then an interval of 15 days (plus or minus) will be used.

Dates associated with the primary endpoint with both month and day of month missing will be imputed as July 1 of the known year provided another date in the same year is to be used in the computation of an interval. If both the month and day of the month are missing and another date in the same year is to be used in the computation of an interval, then the missing date components will be taken as the half the distance from the other date until the beginning or end of the year (depending on expected sequence). If both dates to be used in the computation of an interval are missing both month and day of the month then an interval of 180 days (plus or minus) will be used.

8.2.2. Survival Analysis Method

Imputation of missing date parts will follow the rules previously specified. Competing events for death are "definite" lost to follow-up and withdrawal of consent for continued vital status assessments. Both of these competing events are subject to revision for an individual subject with the passage of time.

The primary analyses will be performed on the ITT set and will analyze subjects according to the randomized arm. The primary analysis will be repeated on the FAS as a sensitivity analysis.

The primary test of each endpoint will consist of a stratified logrank test for the test of the null hypothesis (STRATA <stratification-variable> / TEST=LOGRANK;). See **Section 4.2** for the definition of the stratification variable. The significance of the logrank test will be assessed by referring the one-sided stratified logrank test P_1 to 0.0125, and the comparison will be said to meet statistical criterion if $P_1 < 0.0125$.

Notes:

- These analyses must be stratified: Anderson et al. 2006.
- When there are competing events then censoring competing events provides a between-arm test of the event of interest: Prentice et al. 1978.

The hazard ratio for each comparison will be estimated using stratified (STRATA <stratification-variable>;) proportional hazard regression with the specification of ties handled by the exact method (TIES=EXACT option on MODEL statement). Two-sided 95% confidence interval will be reported in order to maintain consistency with the reporting of estimates from other prostate cancer studies.

8.2.3. Survival Graphs

The survival data will be plotted as either Kaplan-Meier or cumulative incidence distribution estimates depending on the amount of competing events. The cumulative incidence estimates and not the Kaplan-Meier estimates will be plotted if the percent of subjects with competing events in either arm exceeds 3%. When there are competing events then the Kaplan-Meier estimate is not an estimate of the probability free of event: Gooley et al, 1999.

8.2.4. Other Analyses of Survival

The following other analyses, supporting, sensitivity, or exploratory, will be performed:

- Deaths that occur after the cutoff will not be censored.
- Clinical site homogeneity will be assessed by including in the model site grouping multi-level main effect terms and also interaction terms for arm by categorical site grouping.
- The analysis will be re-run with arm by stratum interaction added to the model statement in order to assess arm by stratum interaction. If the 2-sided P value for this multiple degree-of-freedom test of homogeneity is significant at two-sided 0.10, then the stratum-specific hazard ratio estimates and confidence intervals will be computed. If the statistical criterion is not met then the test of homogeneity will be regarded as an exploratory analysis of the data collected. Thus, the outcome of the strata homogeneity test will not affect the interpretation of significance for the primary test, but will affect the way the results are presented or subsequently analyzed.
- Effect modifier analyses will be performed for putatively prognostic factors, including those explored previously in the literature as well as those hypothesized to be predictive in the context of this study. Effect modification will be assessed one factor at a time in stratified models. Each factor will be analyzed as either a natural dichotomy, a dichotomized ordered categorical with cutpoint chosen based on biologic considerations, or a continuous variable dichotomized at the overall median. A separate effect size estimate for arm and its 95% confidence interval will be computed for each level of the dichotomy using contrast statements in SAS.
- The four Kaplan-Meier estimates (arm by factor) for each of the factors analyzed as effect modifiers will be plotted on one graph.
- Regression models with multiple added prognostic covariates will be estimated. The covariates to be added will include those explored previously in the literature as well as those hypothesized to be predictive in the context of this study. Interactions will also be explored for contribution to the model.
- Subjects randomized and not initiating study intervention will be added to the FAS and the primary analyses repeated.
- An analysis of the relationship between dose intensity and outcome by arm will be performed. The main method will be a time-dependent covariate that increases by one for each injection. The interaction between arm and this time-dependent covariate will be the focus of this analysis.

8.3. Secondary Endpoint (Alive Without Event [AWE] at Six Months)

The testing of AWE will have meaning in the context of type I error probability control only if the corresponding primary survival comparison meets its formal statistical criterion. The formal significance levels used for survival analysis will be used for AWE.

AWE has less statistical sensitivity than an analysis based on the closely related and traditional time-to-progression because the timing of events is ignored. Thus, AWE is inherently conservative. This trade-off is acceptable because the trial size provides adequate sensitivity for meaningful differences in the AWE proportions for each comparison. The approximate statistical sensitivity of AWE can be illustrated by assuming a simple comparison of proportions between the arms being compared. Assume a significance level of one-sided 0.0125 and that the control arm proportion is 0.5. Each comparison is based on 800 subjects, 400 per arm. The probability is 90% or greater that this statistical test will be significant if the true proportion difference is 0.126 or greater and the critical difference is 0.080.

The 95% confidence interval on the odds ratio estimate will be computed as the measure of the magnitude of effect.

All analyses of AWE will be conducted on the ITT set and repeated for the FAS as a sensitivity analysis.

8.3.1. AWE Analysis Method

The AWE endpoint will be analyzed using stratified logistic regression. The estimated odds ratio will be greater than unity if the experimental arm has a more favorable outcome.

Using the estimated odds ratio and a reference specification for the proportion of subjects with AWE in the control arm the expected proportion of subjects with AWE for the investigational arm can be projected, and thus the projected between-arm difference in proportions computed.

If the statistical criterion is met then a test for homogeneity of the odds ratio across strata will be performed by adding arm by stratum interaction to the above model. If the P value for this multiple degree-of-freedom test of homogeneity is significant at two-sided 0.10, then the stratum-specific odds ratios and confidence intervals will be computed using the ESTIMATE feature of PROC LOGISTIC. If the statistical criterion is not met then the test of homogeneity will be regarded as an exploratory analysis of the data collected. Thus, the outcome of the homogeneity test will not affect the interpretation of significance for the primary test, but will affect the way the results are presented or subsequently analyzed.

8.3.2. Other Analyses of AWE

The following other analyses, supporting, sensitivity, or exploratory, will be performed:

- Clinical site homogeneity will be assessed by including in the model site grouping multi-level main effect terms and also interaction terms for arm by categorical site grouping.
- The analysis will be re-run with arm by stratum interaction added to the model statement in order to assess arm by stratum interaction. If the 2-sided P value for this multiple

degree-of-freedom test of homogeneity is significant at two-sided 0.10, then the stratum-specific hazard ratio estimates and confidence intervals will be computed. If the statistical criterion is not met then the test of homogeneity will be regarded as an exploratory analysis of the data collected. Thus, the outcome of the strata homogeneity test will not affect the interpretation of significance for the primary test, but will affect the way the results are presented or subsequently analyzed.

- Effect modifier analyses will be performed for putatively prognostic factors, including those explored previously in the literature as well as those hypothesized to be predictive in the context of this study. Effect modification will be assessed one factor at a time in stratified models. Each factor will be analyzed as either a natural dichotomy, a dichotomized ordered categorical with cutpoint chosen based on biologic considerations, or a continuous variable dichotomized at the overall median. A separate effect size estimate for arm and its 95% confidence interval will be computed for each level of the dichotomy using contrast statements in SAS.
- Regression models with multiple added prognostic covariates will be estimated. The covariates to be added will include those explored previously in the literature as well as those hypothesized to be predictive in the context of this study. Interactions will also be explored for contribution to the model.
- Subjects who had not been previously observed to have an event and who miss their six month disease assessments are recoded as AWE = 1 (instead of AWE = 0) for both arms, and then for each arm separately.
- Subjects who had not been observed previously to have an event and who have assessment(s) that are not in the window for the six month assessment are coded according to the closest assessment, for both arms, and then for each arm separately.
- The analysis of progression will be subjected to a tipping point analysis as described in Yan et al. 2009.

8.4. Exploratory Endpoint (Alive Without Event [AWE] at the Six-Month LTFU Visit)

The AWE at the 6 month LTFU visit endpoint will be analyzed using stratified logistic regression. The estimated odds ratio will be greater than unity if the experimental arm has a more favorable outcome.

The 95% confidence interval on the odds ratio estimate will be computed as the measure of the magnitude of effect.

All analyses of AWE at the 6 month LTFU visit will be conducted on the ITT set, excluding subjects not alive at the Week 25 (EOT) visit.

8.5. Tertiary Objectives

The tertiary objectives are exploratory and will be analyzed using statistical modeling. These exploratory analyses will not be conditional on meeting a primary objective.

8.5.1. HLA-A2 Status

Positive HLA-A2 status is defined by a positive serological response to a HLA-A2-specific antibody. The previously completed phase 2 found that HLA-A2 positive subjects had a larger effect size than did HLA-A2 negative subjects. This finding will be confirmed in this study through an effect modification analysis in the primary model (including stratification). The primary proportional hazard regression model will be re-analyzed with the addition of HLA-A2 status (positive vs. negative) and HLA-A2 status by treatment interaction terms. The criterion for significant effect modification will be a P value of 0.1 or less for the coefficient of the HLA-A2 status by treatment interaction term. If there is significant effect modification then the hazard ratios specific to negative and positive HLA-A2 status will be estimated along with additional supporting analyses such as Kaplan-Meier estimate graphs.

8.5.2. Absence of Progression at 12 Months

To model the likelihood of remaining radiologically event-free at 12 months as compared to scans at 6 months is of interest. Subjects may become secondarily stable, allowing for time for the vaccine to induce an immune response, and then establish a new baseline at 6 months. The between-arm comparison of this outcome is not a comparison free of the confounding of post-randomization events that have the potential of being arm-dependent. Regression modeling of being free of progression, which is AWE at 12 months, with covariates for baseline prognostic factors and AWE at 6 months, will be conducted.

8.5.3. Post-Progression Anti-Cancer Therapies

Because the primary endpoint is survival and subjects can have access to other anticancer therapies following progression there is the possibility that post-progression anticancer therapies (PPACTs) could provide an alternative explanation for any observed survival differences. Analyses for assessing the possibility of PPACT effects are necessarily open-ended because the challenge is to find an alternative explanation for an observed outcome from a pre-planned analysis. Thus, such analyses are necessarily data-directed because they are challenge analyses, leaving the assessment of meaning in the context of being data-directed and the multiplicity of analyses to be assessed through consideration of clinical and biological considerations.

The following classes of analyses are anticipated at this time:

- Compare the Kaplan-Meier estimates from each arm to the Kaplan-Meier estimate where subjects are censored at the time of initiation of a PPACT.
- Model the use of PPACTs as time-dependent covariates added to the primary survival models (FAS, arms as randomized). In addition, baseline covariates of prognostic importance will be added to these models in order to attempt adjustment for prognosis factors in the selection of subjects getting PPACTs (confounding between subject attributes and selection for PPACTs).

8.5.4. Immune Responses

PROSTVAC has been found to induce various immune responses. The arms will be compared with respect to previously identified immune responses and other immune responses that may not have been previously identified. The prognostic value of immune responses will be assessed.

8.5.5. Circulating Tumor Cells

Technology for assessing for CTCs has become more available and reliable, however limited information exists on CTC counts in the target population. Modeling will be used to assess the prognostic value of baseline CTC measures and other prognostic measures on outcome.

8.5.6. Tumor Regression and Growth Rate

Tumor regression and growth rates will be derived for each subject with sufficient PSA data per the method described in Stein, et al, 2010. Linear regressions to evaluate the relationship between the growth rate constant, g, or other parameters and survival will be performed. Regression rate constants, d, and growth rate constants, g, will be presented in the data listings for each subject.

8.5.7. Overall Prognostic Model

Additional data-driven analyses may be performed at the discretion of the sponsor. The tertiary objectives above focus on specific classes of factors that are potential explanatory of survival. There will be an attempt to model all of these factors, including arm, potential baseline prognostic factors collected at screening, including baseline subject attributes and circulating tumor cells, immune responses, the use of post-progression anti-cancer therapies, and tumor regression and growth rates for their relationship to survival.

9. Safety Analysis

All safety analyses will be conducted on the FAS. All summaries in the safety analysis will be presented by treatment arm and overall.

9.1. Adverse Events

Adverse event terms recorded on the eCRFs will be coded to PTs using version 11.0 or later of the Medical Dictionary for Regulatory Activities (MedDRA). Toxicity grade will be defined according to the version 4.0 or later of the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE).

A treatment emergent adverse event (TEAE) is defined as any AE with an onset date on or after date of first dose of study product, or any ongoing event on the date of first dose that increases in CTCAE grade after date of first dose. Only TEAEs with an onset date before date of last dose + 28 days will be tabulated in summary tables. Adverse events that begin prior to first dose will be captured in the medical history.

Furthermore each TEAE will be classified based on the period in which it started. Primary period TEAEs will be defined as TEAEs with onset date on or after the date of the first dose of study product and before the date of the second dose of study product. Booster period TEAEs will be defined as TEAEs with onset date on or after the date of the second dose of study product.

For the purpose of calculating treatment emergence, period, and inclusion in summary tables, incomplete onset dates will be imputed as detailed in **Appendix C** of this document.

9.1.1. Incidence of Adverse Events

The frequency and percentage of subjects with TEAEs will be tabulated for overall incidence by system organ class (SOC) and PT for primary period TEAEs, booster period TEAEs and overall.

TEAEs will also be summarized by treatment arm and overall for the highest CTCAE grade and highest relationship to study treatment within each subject. Related TEAEs, serious TEAEs, related serious TEAEs, and TEAEs resulting in study product discontinuation will be similarly summarized. Additionally, the incidence of TEAEs and related TEAEs will be summarized by PT in descending order of total frequency. The incidence of TEAEs coded as injection site reactions will be summarized by type of injection (primary or booster) and by arm.

At each level of summarization, a subject will be counted only once for each AE he/she experiences within that level.
Data listings of all AEs and SAEs will be provided. Adverse events presented in the listings will be those recorded from the first dose date through the end of follow-up.

9.1.2. Survival Status and Death

Subject survival status will be summarized in a table using the FAS. Furthermore, all subject deaths during this study will be presented in a listing. The listing will provide all relevant eCRF data pertaining to each subject death.

9.2. Clinical Laboratory Evaluations

The following laboratory parameters will be summarized:

- Serum Chemistry - alanine aminotransferase (ALT), alkaline phosphatase (ALP), aspartate aminotransferase (AST), bicarbonate, blood urea nitrogen (BUN), calcium, chloride, creatinine, glucose, lactate dehydrogenase (LDH), potassium, sodium, testosterone, and total bilirubin.
- Hematology - basophils, eosinophils, hematocrit (HCT), hemoglobin (HGB), lymphocytes, monocytes, neutrophils, platelets, red blood cells (RBC) and white blood cells (WBC).
- Urinalysis - protein, glucose, ketones, blood, leukocyte esterase, specific gravity, and pH
- Coagulation - PT, PTT, INR

- Virology – HIV Viral Load, Hepatitis B Surface Antigen (HBSAG), Hepatitis C Viral Load (HCV)
- Additional - Prostate Specific Antigen (PSA), Prostatic Acid Phosphatase (PAP), C-Reactive Protein (CRP)

Change from baseline tables will be presented for laboratory tests with numeric results. Mean change from baseline to each scheduled post-baseline visit, last visit with available assessment, and minimum and maximum post-baseline assessments will be presented for each test. Shift from baseline tables by CTCAE v4.0 (or higher) grades (where applicable) and by high/low flags (where CTCAE grades are not defined) will be presented to show the number and percentage of subjects in each cell. All scheduled post-baseline visits, last visit with available assessment, and worst shift overall will be included in shift tables. Laboratory tests (calcium, glucose and sodium) that have both high and low abnormalities will be summarized separately by CTCAE v4.0 (or higher) grades for each direction of abnormality (e.g. hypercalcemia vs. hypocalcemia).

The mean and standard deviation over time of selected analytes will be presented in a figure for the FAS.

All collected laboratory test results will be presented in the data listings.

9.3. Cardiac Markers

Cardiac markers, including CK, CK-MB, troponin-t, and troponin-i, will be collected at screening and presented in the data listings.

9.4. Vital Sign Measurements

Vital sign measurements data will be summarized for the FAS. Sitting systolic and diastolic blood pressure, temperature, respiratory rate, pulse rate, and weight will be summarized for actual result, change from baseline to each scheduled post-baseline visit, last visit with available assessment, minimum and maximum post-baseline assessments. Vital sign measurements data will also be presented in a listing. The mean and standard deviation over time for blood pressure will be plotted in a figure.

9.5. Physical Examination

Physical examination data will be presented in a listing.

9.6. Electrocardiogram

A shift from baseline table by overall Electrocardiogram (ECG) interpretation will be presented to show the number and percentage of subjects in each cell.

ECG data will be presented in a listing.

9.7. Immune Monitoring

Immune Monitoring data will be presented in the data listings.

10. Quality of Life

Subject-reported quality-of-life (QOL) data will be presented in the data listings.

11. Radiology

Radiological assessments will be performed at Screening, Week 13, Week 25 EOT, and at the 6-month long term follow-up visit.

Radiology data will be presented in the data listings.

12. Interim Analysis

This section describes the interim analysis plan (IAP) to be executed by the independent statistician and assessed by the DMC. The specific IAP results are closed to the sponsor unless one of the criteria are met, and even then the DMC will use its discretion regarding the communications of results to the sponsor, with patient safety balanced against benefit being the primary concern. This IAP consists of the ability to detect early evidence of efficacy and also futility.
The statistical computations used to devise this IAP take into account the possibility of a delayed effect of the same magnitude as used to plan the trial size but the statistical tests will ignore the possibility of delayed effect because there cannot be certainty that a delayed effect exists. Stated another way, the statistical tests to be done as specified in the IAP are optimal in cases where the hazard ratio is proportional; the alternative is intractable because it would be to use parametric statistics in a situation where the underlying form of the failure-time distribution is not known. Thus, if the possibility of a delayed effect were not taken into account and it actually exists then the IAP would have lower power to detect early efficacy than anticipated because early events would not be discriminatory for efficacy, and a conclusion of futility would have increased power because early events would be give a false signal about lack of efficacy. By contrast, if the possibility of delayed effect were taken into account when a delayed effect does not exist then the IAP will have greater power than anticipated to detect early efficacy whereas futility would have diminished power than anticipated.

The IAP is designed so that interim analyses for a comparison are to be executed at 40%, 60%, and 80% information times, corresponding to 214, 321, and 427 deaths, with the final analysis at 534 deaths as previously discussed.

The one-sided significance criteria to be used for early detection of efficacy analyses are 0.000135, 0.00147, 0.00501, respectively, with the adjusted final analysis to use 0.01063. These significance levels were computed using SAS PROC SEQDESIGN as one-sided O'Brien-Fleming boundaries with an overall one-sided type I error probability of 0.0125. Following is the SAS code used to compute these intervals.

```
proc seqdesign boundaryscale=pvalue;
  design
    nstages=4
    info=cum(0.4 0.6 0.8 1)
    method=obf
    alt=upper
    stop=reject
    alpha=0.0125
  ;
run;
```

The early detection of efficacy is to use the same statistical testing methodology as specified for the primary analysis. The statistical characteristics for the interim analysis event-driven time-points and significance levels are computed using simulations. The simulation methodology for delayed effect is described in **Appendix A**. The following table shows the statistical operating characteristics of this plan based on 10,000 simulation replicates. For example, when the alternative hypothesis is true (average hazard ratio = 0.68) then at the second interim analysis is projected for 3.2 years following the start of randomization and the estimated cumulative probability of rejecting the null hypothesis is 0.1571 (with confidence interval as shown). Also shown is the largest hazard ratio found for which the null hypothesis was rejected (rej HR), showing that at this interim analysis a hazard ratio less than or equal to 0.716 will pass the criterion for early efficacy (an estimate of the critical hazard ratio). Please note that these simulations are consistent with the targeted overall one-sided type I error of 0.0125: for analysis 4 ("#" column) the proportion of trials rejected ("rejected" column) under the null hypothesis (row with "null" in the "hyp" column) is 0.0127, and is close enough to 0.0125 considering computations were done using simulation.

Statistical Operating Characteristics of the Interim Analysis Plan

Simulation estimates (cumulative, # replications = 10000):

#	events	hyp	time	test	rejected	95% CI	rej HR
1	214						
		null	2.5	futile	0.0771	0.0719 to 0.0825	>=1.212
				superior	0.0000	0.0000 to 0.0004	<=0.000
		alt	2.5	futile	0.0054	0.0041 to 0.0070	>=1.216
				superior	0.0065	0.0050 to 0.0083	<=0.603
2	321						
		null	3.0	futile	0.2297	0.2215 to 0.2381	>=1.090
				superior	0.0013	0.0007 to 0.0022	<=0.712
		alt	3.2	futile	0.0071	0.0055 to 0.0089	>=1.093
				superior	0.1571	0.1500 to 0.1644	<=0.716
3	427						
		null	3.7	futile	0.4224	0.4127 to 0.4322	>=1.023
				superior	0.0045	0.0033 to 0.0060	<=0.779
		alt	4.0	futile	0.0078	0.0062 to 0.0097	>=1.032
				superior	0.5417	0.5319 to 0.5515	<=0.779
4	534						
		null	4.6	superior	0.0127	0.0106 to 0.0151	<=0.819
		alt	5.0	superior	0.8388	0.8314 to 0.8460	<=0.819

Notes:

- Delayed effect: HR = 1 for 4 months, four month linear ramp to HR = 0.68, with HR = 0.68 going forward.
- Overall $\alpha = 0.025$ (0.0125 for each comparison)
- ~85% target power
- The total accrual period is 2.5 years with a 19 month linear ramp to constant accrual (800 in each comparison, 400 per arm)
- Control arm median = 22 months

The significance level to be used for detection of futility at each interim analysis is 0.00001. This futility significance level was chosen in order to provide a balance between the sought for estimated probability of detecting futility and protecting against falsely declaring futility. A preceding table also shows the statistical operating characteristics of futility detection. For example, when the null hypothesis is true (hazard ratio = 1) then at the second interim analysis is projected for 3.0 years following the start of randomization and the estimated cumulative probability of rejecting the futility null hypothesis is 0.2297 (with confidence interval as shown). Also shown is the smallest hazard ratio found for which the futility null hypothesis was rejected (rej HR), showing that at this interim analysis a hazard ratio greater than or equal to 1.090 will pass the criterion for futility (an estimate of the futility critical hazard ratio).
The futility null hypothesis is H_0: $\Delta \leq 0.68$ and the futility alternative hypothesis is H_A: $\Delta > 0.68$, where Δ is the hazard ratio. The testing of the futility null hypothesis using SAS is computed as follows. The futility null hypothesis can be rewritten using the natural logarithm as

H_0: $\ln(\Delta) \leq \varphi$, or H_0: $\beta_{arm} \leq \varphi$), or H_0: $\beta_{arm} - \varphi \leq 0$.

where $\varphi = \ln(0.68)$ and β_{arm} is the logarithm of the hazard ratio associated with the experimental arm. The futility null hypothesis is tested using SAS PHREG using the OFFSET option.

The possibility of there being no delay of effect can be assessed through simulation where the delay of effect is eliminated. The following table shows the statistical operating characteristics of the IAP when no delay of effect is applicable. When there is no delay of effect there is considerably more power to detect early efficacy (a similar outcome was observed for the trial planning computations).

Statistical Operating Characteristics of the Interim Analysis Plan for No Delay of Effect

Simulation estimates (cumulative, # replications = 10000):

#	events	hyp	time	test	rejected	95% CI	rej HR
1	214						
		null	2.5	futile	0.0780	0.0728 to 0.0834	>=1.213
				superior	0.0002	0.0000 to 0.0007	<=0.600
		alt	2.6	futile	0.0000	0.0000 to 0.0004	NA
				superior	0.1902	0.1825 to 0.1980	<=0.605
2	321						
		null	3.0	futile	0.2249	0.2167 to 0.2332	>=1.090
				superior	0.0008	0.0003 to 0.0016	<=0.716
		alt	3.3	futile	0.0000	0.0000 to 0.0004	NA
				superior	0.6745	0.6652 to 0.6837	<=0.717
3	427						
		null	3.7	futile	0.4232	0.4135 to 0.4330	>=1.024
				superior	0.0051	0.0038 to 0.0067	<=0.778
		alt	4.1	futile	0.0002	0.0000 to 0.0007	>=1.043
				superior	0.9218	0.9164 to 0.9270	<=0.779
4	534						
		null	4.6	superior	0.0119	0.0099 to 0.0142	<=0.819
		alt	5.1	superior	0.9854	0.9829 to 0.9877	<=0.819

Notes:

- No delayed of effect: HR = 0.68 for all time
- Overall $\alpha = 0.025$ (0.0125 for each comparison
- ~85% target power
- The total accrual period is 2.5 years with a 19 month linear ramp to constant accrual (800 in each comparison, 400 per arm)
- Control arm median = 22 months

There are IAP issues that require special consideration because of the nature of this study: The events needed for each of the two comparisons may not occur at the same time. The IAP for an interim analysis and the final analysis is to be implemented only when the required numbers of events have been reached for both comparisons.

More than the required number of events may be applicable at the time of an interim or final analysis. The definitive analysis will be based on backing up the data until the time point at which the required number of deaths is first applicable. An analysis using all data available will also be executed as a supplemental analysis.

The DMC will be required to use its discretion regarding the action to be taken in case one comparison meets criterion at an interim analysis whereas the other comparison does not. The issue is whether an action for the comparison meeting criterion would adversely affect the ability to obtain a definitive answer to the comparison not meeting criterion.

13. References

Crowley J and Ankerst DP eds. *Anderson GL, LeBlanc M, Liu PY, Crowley J: On the use of covariates in randomization and analysis in clinical trials*. Handbook of Statistics in Clinical Oncology. 2nd ed. Boca Raton: CRC Press. 2006.

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

Prentice RL, Kalbfleisch JD, Peterson A, Flournoy N, Farewell V, Breslow N. "The analysis of failure times in the presence of competing risks". *Biometrics* 1978; 34: 541-554.

Shih JH. "Sample size calculation for complex clinical trials with survival endpoints." *Controlled Clinical Trials* 1995; 16:395-407.

Stein WD, Gulley JL, Schlom J, Madan RA, Dahut W, Figg WD, Ning Y, Arlen PM, Price D, Bates SE, Fojo T. "Tumor Regression and Growth Rates Determined in Five Intramural NCI Prostate Cancer Trials: The Growth Rate Constant as an Indicator of Therapeutic Efficacy". *Clin. Cancer. Res.* 2011; 17: 907-917.

Yan X, Lee S, Li N. "Missing Data Handling Methods in Medical Device Clinical Trials". *J. Biopharm. Stat.* 2009; 19: 1085-1098.

14. Appendix A: Delayed Effect Simulation Methodology

The trial size and interim analyses were planned using clinical trials simulation software developed by Brent A. Blumenstein, PhD, the statistical consultant providing the statistical design of this study. This software is quite general and has been under development since the 1980s. Validation has included comparison of results from other software used for trial planning, including SAS and East. The software used provides results that are "close enough" for the cases that can be compared and considering that simulation results are being compared to formulaic methods.

The reason this software is used instead of other more commonly available methods is because this software has the ability to simulate trials with a specific type of departure from proportional hazards, specifically the case where the target hazard ratio is not initially applicable but becomes applicable later in time. Neither SAS nor East has this capability, for example, and base their computations on assuming proportional hazards. This appendix subsequently discusses the details of the methodology implemented in the clinical trials simulation software used by BNIT.

The type of departure from proportional hazards can be generically identified as a later separation of Kaplan-Meier estimates, and there is conjecture that immunotherapies are subject to delayed effect consistent with what is seen when Kaplan-Meier estimates have delayed separation.

The observation of delayed separation of the BNIT phase 2 trial survival Kaplan-Meier estimates was the reason the phase 3 trial was planned with delayed effect and therefore the simulation software implementing this type of departure from proportional hazards was used. If the delayed effect conjecture was not used to plan the BNIT phase 3 trial then the trial could be seriously underpowered and the interim analysis plan could lead to meeting futility criterion earlier than valid. The clinical trials simulation software criteria generated for early detection of superiority also prevents inappropriate triggering of superiority stopping rules.

There is limited ability to validate the results from this software because the methodology for implementing delayed effect is relatively new. Delayed effect validation has been implemented in a variety of ways, including assessing whether the gradual implementation of delayed separation gives results consistent with expectations and comparison to some other implementations of delayed separation. For example, ECOG used a sum of exponentials survival distribution and numerical integration to plan some leukemia trials where delayed separation of Kaplan-Meier estimates was apparent. The R code implementing this numerical integration trial planning methodology was provided and the clinical trials simulation software results implementing the intended delay were compared and the results were favorable within the limits of the fundamental differences in methods. The results from the clinical trial simulation software have also been compared with the same comparability conclusion to software implementing a discrete chain Markov method as described by Shih [Shih JH. Sample size calculation for complex clinical trials with survival endpoints. Controlled Clinical Trials 1995; 16:395-407.].

Following is a description of the simulation methodology used to evaluate the statistical operating characteristics of trials where the intervention under study is subject to a delay of effect. (The text that follows is based on a methodological manuscript submitted for publication.)

The reasons for a delayed effect include: (1) Randomization of subjects with advanced disease for which experimental intervention is "too late". (2) The experimental intervention has a delayed onset of benefit. One or both of these mechanisms could apply.
Assume a two-arm randomized clinical trial and that individuals entering this trial are to be randomized to one of two groups designated as arms 1 and 2, where arm 1 is to receive control intervention (perhaps a placebo) and arm 2 is to receive experimental intervention. Assume that the outcome of interest is time to a failure event or time to the last date when this event has yet to be observed. Assume it is hypothesized that the experimental intervention induces a distribution of event times that is more favorable, that is, shifted to the right, as compared to the distribution of event times associated with the control intervention. Assume competing events are either nonexistent or negligible.

Trials of this type are typically planned by computing the required number of events under the assumption of proportional hazard rates, specific type I and II error probabilities, and a hypothesized clinically consequential ratio of the arm 2 (experimental) hazard rate to the arm 1 (control) hazard rate. Let h_A be the hypothesized consequential hazard ratio.

Let the hazard ratio function $h(t) = \lambda_1(t)/ \lambda_2(t)$ where $\lambda_1(t)$ and $\lambda_2(t)$ are hazard rate functions for arm 1 and 2, respectively. The proportional hazard rates assumption means that $h(t) = \lambda_1(t)/ \lambda_2(t) = h$ (a constant).

After computing the required number of events, projections for the timing of analyses are usually made based on assuming exponential distributions for the arms with hazard rates $\lambda_1(t) = \lambda_1$ and $\lambda_2(t) = \lambda_2 = h_A\lambda_1$. Thus, when there is no delayed effect the survival functions for arm 1 and arm 2 are usually specified to be

$$F_1(T \leq t) = \exp(-\lambda_1 t)$$
$$F_2(T \leq t) = \exp(-\lambda_2 t) = \exp(-h_A\lambda_1 t).$$

The delayed effect simulation methodology described here is based on assuming that the control arm survival distribution is exponential. (It is possible to generalize this methodology to other types of control arm survival distributions though the use of the exponential distribution is usual.)

Now, assume that the experimental intervention has a lag in realization of effect so that the arm 2 hazard rate is initially λ_1 and only later becomes $h_A\lambda_1$. This delay of effect assumes the arm 2 hazard rate function is

$$\lambda_2(t) = (\lambda_1 - \lambda_2)g(t) + \lambda_2, (t = 0),$$

where g(t) is a function with the following properties: $g(0) = 1$, g(t) is non-increasing, and either $g(t) = 0$ for some t' where $t \geq t' > 0$ or $g(t) \rightarrow 0$ as $t \rightarrow \infty$. This implementation of delayed effect specifies that ultimately the hazard ratio becomes proportional or approaches proportionality.

In order to do simulations under delayed effect the survival function $F_2(t)$ must be derived. This survival function is obtained using the relationship

$$F_2(t) = \exp(-H_2(t))$$

where H2(t) is the integral of $\lambda_2(t)$ from 0 to t (integrated hazard function). Depending on the function g(t) the integral $H_2(t)$ may have an analytic form; if not then numerical integration can be used.

A segmented linear form of g(t) will now be described. (Other forms of g(t) are possible.) The segmented linear form of g(t) has two parameters, d and r, where $d \geq 0$ and $r > 0$. The parameter d is the delay in the start of the effect of the experimental intervention and r is the time it takes to realize full effect from the start of realization of effect. The function g(t) is specified linear and descending between d and d + r.. Specifically, g(t) is defined as follows

$g(t) = 1$ for $t < d$
$g(t) = (d + r – t)/r$ for $d \leq t < d + r$,
$g(t) = 0$ for $t \geq d + r$.

It follows that

$H_2(t) = (\lambda_1 - \lambda_2)t + \lambda_2 t = \lambda_1 t$ for $t < d$
$H_2(t) = (\lambda_1 - \lambda_2)\{d + [(d + r)(t – d) –(t^2 – d^2)/2]/r\} + \lambda_2 t$ for $d \leq t < d + r$,
$H_2(t) = (\lambda_1 - \lambda_2)(d + r/2) + \lambda_2 t$ for $t \geq d + r$.

Pseudo random survival times from F_2 can be generated by tabling its inverse and using a uniform pseudo random generator.

Thus, if F_1 is exponential (as usual) and F_2 is defined as above then pseudo random samples of survival times from F_1 and F_2 can be generated for specified values of λ_1, λ_2, d and r. These survival times can be converted to clinical trial data based on specification of accrual distribution and follow-up intervals.

The consequence of delayed effect is a reduction in projected statistical power relative to no delayed effect. Specifically, events occurring prior to the time of realization of the specified effect h_A will be less differentiating than events that occur after full realization of the specified effect h_A, and therefore only the events occurring after the time of realization of effect fully contribute to statistical power. Thus, if a trial is planned without taking into account delayed effect then the projected power of the trial will be overestimated if delayed effect is present in the data.

Statistical operating characteristics for a specific set of trial specifications, including parameters specifying a delayed effect, are estimated from replicating the generation trial data for the control arm and the experimental arm, that is, from replicates of the simulated trial data. Thus, for specified significance criteria the statistical operating characteristics of the test of the null

hypothesis for superiority of the experimental intervention can be estimated using the logrank test on the simulated trial replicates. The following statistical operating characteristic estimates are most often of interest:

- One-sided type 1 error probability: The proportion of replicates for which the test of the null hypothesis is rejected using the one-sided logrank test and when the experimental arm data are generated with $\lambda_2 = \lambda_1$.
- Power: The proportion of replicates for which the test of the null hypothesis is rejected using the one-sided logrank test and when the experimental arm data are generated with $\lambda_2 = h_A\lambda_1$.

These estimates also can be made for interim analyses based on event time or calendar time, and the cumulative estimates (trial-wise) are those that take into whether and when specific trial replicates meet interim analysis criteria.

The statistical operating characteristics of futility tests can also be estimated using simulation. A common approach to futility testing is based on specifying an additional parameter Ω, usually Ω is the hazard ratio specified in the specific superiority alternative hypothesis, and then testing as the futility null hypothesis that $\lambda_2 = \Omega\lambda_1$ against the futility alternative hypothesis that $\lambda_2 > \Omega\lambda_1$. Evidence against the futility null hypothesis, that is, rejection of the futility null hypothesis, provides a basis for concluding that futility exists. The futility null hypothesis can be tested using a modification of the logrank test or proportional hazard regression where an offset corresponding to Ω is used to set up the futility null hypothesis.

The futility testing scheme described is applicable only when the proportional hazards assumption is tenable, however. Specifically, when delayed effect is applicable then the futility null hypothesis survival distributions will not have proportional hazards. The consequence will be an increase in the likelihood of falsely concluding futility because of the interaction between delayed effect and the offset correction in the statistical test.

Therefore a different approach must be used for futility testing when delayed effect is applicable. The chosen method is to modify the logrank test by using zero weights for event times prior to a pre-specified time T_{eff}, a time specified to be after full effect or nearly full effect is applicable. Specifically, the two-sample logrank statistic is one of a class of linear rank statistics that can be written as

$$\Sigma w_i(d_{i2} - n_{i2}d_i/n_i)$$

where the sum is over each distinct time t_i at which one or more events occur, d_{i2} is the number of arm 2 events at t_i, n_{i2} is the number at risk at t_i, d_i is the total number of deaths at t_i, n_i is the total number at risk at t_i, and w_i is a weight. For example, for the logrank test $w_i = 1$ for all i, and for the Wilcoxon test $w_i = n_i$. The modified logrank statistic therefore has $w_i = 0$ for all $t_i < T_{eff}$, and $w_i = 1$ otherwise. The effect is to focus the futility testing to times after the delay of effect so that the likelihood of falsely concluding futility is controlled but with a reduction in sensitivity for futility.

The operating characteristics of the futility testing can be assessed by generating experimental arm data replicates using F_1 and F_2 (given previously). The logrank test with the offset Ω and modifications of weights using T_{eff} as specified can then be used to obtain the following estimates statistical operating characteristics for futility testing:

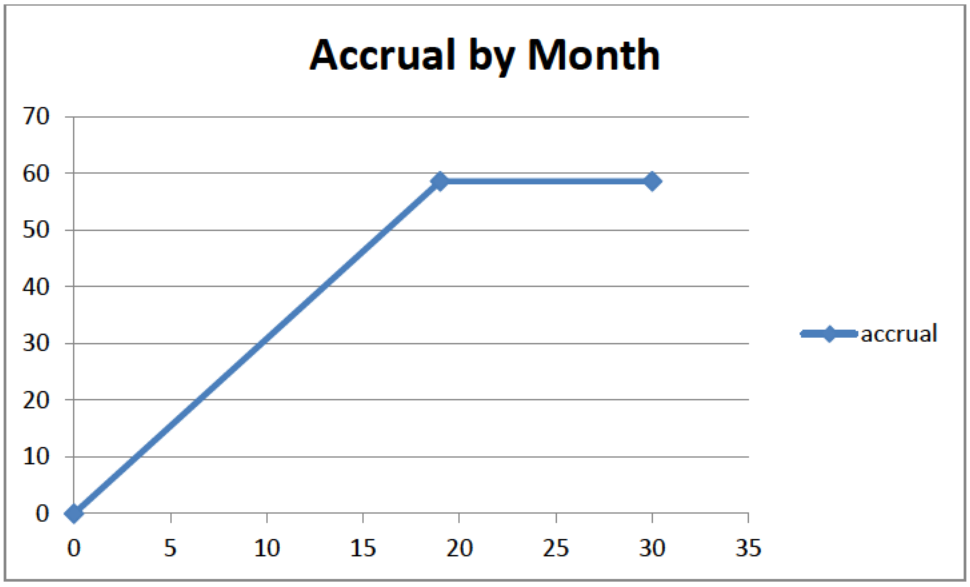
- False futility probability (futility test type I error probability): The proportion of replicates for which the test of the futility null hypothesis is rejected for a specified level of significance. The futility testing will be valid if this proportion is close to the significance level used in the testing of the futility null hypothesis.
- Probability of futility conclusion conditional on specifications of truth: The proportion of replicates for which the test of the futility null hypothesis is rejected. Various values of truth (h_A in the specification of F_2) can be used to assess the likelihood of concluding futility over a range of versions of truth.

A close equivalent to the logrank test described above for futility testing (with an offset and weighting adjusted to T_{eff}) can be implemented as a proportional hazard regression model with a time-dependent covariate for arm. This model has the arm variable as zero prior to time T_{eff} and zero or one for times equal or greater than T_{eff} for the control arm and experimental arms, respectively.

15. Appendix B: Linear Ramp and Accrual Rate

The simulation software used to plan the trial requires specification of the number of subjects to be accrued and the period of time over which accrual is to take place (as is usual with trial planning software). These simple specifications would generate an accrual distribution that is uniform, that is, a constant accrual rate. But since it is unrealistic to assume uniform accrual the software also allows specification of an arbitrary accrual distribution through the specification of intervals and accrual in each interval. An additional feature, and the one used in these computations, allows specification of a linear ramp time for the accrual distribution from which the accrual distribution is generated. The linear ramp time must be less than or equal to the total accrual time. Given the linear ramp time the accrual distribution is generated by assuming zero accrual at time zero and a constant accrual rate increase from time zero through the time specified by the linear ramp time, after which the accrual rate is uniform from the linear ramp time until the end of accrual. The rate of accrual increase during the ramp time and the constant accrual rate after the linear ramp time are derived from the total accrual, the total accrual time, and the linear ramp time. For example, for the computations done in this SAP the following specifications were used: total accrual time of 30 months for 1200 patients and with a linear ramp time of 19 months. Given these three specifications it was computed that the accrual distribution (by month) would have 645 subjects to be accrued after month 19 (uniform accrual rate 645/11 = 58.6 per month for the last 11 months), and that the other 555 subjects were to be accrued in linearly increasing monthly intervals from the start of the trial through month 19 (accrual rate increase of 3.08 per month). This characterization of the approximate accrual distribution was specified as an approximation to the early accrual experience in the ongoing phase 3 trial, and the projections going forward. **Figure 1** illustrates this accrual distribution.

Figure 1: Accrual Distribution



16. Appendix C: Imputation Algorithm for Partial and Missing Dates

Adverse Event

If onset date is completely missing, onset date is set to date of first dose.

If year is present and month and day are missing:
- If year = year of first dose, then set month and day to month and day of first dose.
- If year < year of first dose, then set month and day to December 31.
- If year > year of first dose, then set month and day to January 1.

If year and day are present and month is missing:
- If year = year of first dose, then set month to month of first dose.
- If year < year of first dose, then set month to December.
- If year > year of first dose, then set month to January.

If month and year are present and day is missing:
- If year=year of first dose and
  - if month = month of first dose, then set day to day of first dose
  - if month < month of first dose, then set day to last day of month
  - if month > month of first dose, then set day to first day of month
- if year < year of first dose, then set day to last day of month
- if year > year of first dose, then set day to first day of month

For all other cases, set onset date to date of first dose

Concomitant Medications

Start Date:

If start date is completely missing and end date is not before first dose, then the medication will be classified as both prior and concomitant. Else if start date is completely missing and end date is before first dose, then the medication will be classified as prior.

If (year is present and month and day are missing) or (year and day are present and month is missing), then set month and day to January 1.
If year and month are present and day is missing, then set day to first day of month.

End Date:

If end date is completely missing and start date is not after first dose, then the medication will be classified as both prior and concomitant. Else if end date is completely missing and start date is after first dose, then the medication will be classified as concomitant.

If (year is present and month and day are missing) or (year and day are present and month is missing), then set month and day to December 31.
If year and month are present and day is missing, then set day to last day of the month.

Note that if both start and end dates are missing, then the medication will be classified as prior and concomitant.

17. Appendix D: List of Tables, Listings, and Figures