Title: A Phase 3, Randomized, Double-Blind, Multicenter Study Comparing Oral MLN9708 Plus Lenalidomide and Dexamethasone Versus Placebo Plus Lenalidomide and Dexamethasone in Adult Patients With Newly Diagnosed Multiple Myeloma

NCT Number: NCT01850524

SAP Approve Date: 14 January 2020

Certain information within this Statistical Analysis Plan has been redacted (ie, specific content is masked irreversibly from view with a black/blue bar) to protect either personally identifiable (PPD) information or company confidential information (CCI).

This may include, but is not limited to, redaction of the following:

- Named persons or organizations associated with the study.
- Proprietary information, such as scales or coding systems, which are considered confidential information under prior agreements with license holder.
- Other information as needed to protect confidentiality of Takeda or partners, personal information, or to otherwise protect the integrity of the clinical study.

STATISTICAL ANALYSIS PLAN

A Phase 3, Randomized, Double-Blind, Multicenter Study Comparing Oral MLN9708 Plus Lenalidomide and Dexamethasone Versus Placebo Plus Lenalidomide and Dexamethasone in Adult Patients With Newly Diagnosed Multiple Myeloma

Protocol #: C16014

SAP History

Original	16 March 2018
Amendment 1	14 January 2020

Approved by:

Note: If this document was approved electronically, then the electronic approval signatures may be found at the end of the document.

PPD

_____
Date

_____
Date

_____
Date

_____
Date

Rationale for Amendment 1

This document describes the changes in reference to the statistical analysis plan (SAP) incorporating Amendment No. 1. The primary reason for this amendment is to modify the SAP to ensure timely analysis of the primary endpoint, progression-free survival (PFS), in light of the slower than expected PFS event rate over the past year. The second interim analysis (IA) – the final analysis for PFS – will now take place when approximately 370 PFS events have been observed. Power remains sufficient at 92%.

Minor grammatical, editorial, formatting, and administrative changes are included for clarification purposes only.

Changes in Amendment 1

1. Update statistical procedures to modify the number of events for the final PFS analysis.
2. Clarify the statistical boundary for PFS at the second IA.
3. CCI [redacted]
4. Update list of covariates in the adjustment of overall survival analysis for potential confounding effects by subsequent therapies after patients discontinue study treatment.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND GLOSSARY OF TERMS

Abbreviation	Term
AE	adverse event
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANC	absolute neutrophil count
ASCO	American Society of Clinical Oncology
AST	aspartate aminotransferase
ASCT	autologous stem cell transplant
BM	bone marrow
BSA	body surface area
CBC	complete blood count
CFR	Code of Federal Regulations
CL	clearance, IV dosing
CL_P	plasma clearance
CL_{Total}	total clearance
C_{max}	single-dose maximum (peak) concentration
CO_2	carbon dioxide
CR	complete response
CT	computed tomography
CTCAE	Common Toxicity Criteria for Adverse Events
C_{trough}	single-dose end of dosing interval (trough) concentration
CV	coefficient of variation
CYP	cytochrome P_{450}
DDI	drug-drug interaction(s)
DLT	dose-limiting toxicity
DNA	deoxyribonucleic acid
DOR	duration of response
ECG	electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	electronic case report form
EDC	electronic data capture
EOS	End of Study (visit)
EOT	End of Treatment (visit)

Abbreviation	Term
EU	European Union
FA	final analysis
FDA	United States Food and Drug Administration
GCP	Good Clinical Practice
GI	gastrointestinal
GLP	Good Laboratory Practices
GM-CSF	granulocyte macrophage-colony stimulating factor
GMP	Good Manufacturing Practice
Hb	hemoglobin
HU	health utilization
IB	Investigator's Brochure
ICF	informed consent form
ICH	International Conference on Harmonisation
IDMC	Independent Data Monitoring Committee
IEC	independent ethics committee
CCI	[redacted]
IMWG	International Myeloma Working Group
IRB	institutional review board
IRC	independent review committee
ITT	intent-to-treat
IV	intravenous; intravenously
IVRS	interactive voice response system
KPS	Karnofsky Performance Status
LFT	liver function test(s)
MedDRA	Medical Dictionary for Regulatory Activities
MID	minimally important difference
Millennium	Millennium Pharmaceuticals, Inc., and its affiliates
MM	multiple myeloma
MRI	magnetic resonance imaging
MRU	medical resource utilization
MTD	maximum tolerated dose
NCCN	National Comprehensive Cancer Network
NCI	National Cancer Institute
NCI CTCAE	National Cancer Institute Common Terminology Criteria for Adverse Events

Abbreviation	Term
NDMM	Newly diagnosed multiple myeloma
CCI	[redacted]
ORR	overall response rate
OS	Overall survival
PBMC	peripheral blood mononuclear cell
PD	progressive disease (disease progression)
PFS	Progression-free survival
PK	pharmacokinetic(s)
PO	*per os*; by mouth (orally)
PR	partial remission ***or*** partial response
PRO	patient-reported outcome(s)
PSA	prostate-specific antigen
CCI	[redacted]
QOL	quality of life
QTc	rate-corrected QT interval (millisec) of electrocardiograph
RBC	red blood cell
RECIST	Response Evaluation Criteria in Solid Tumors
SAE	serious adverse event
SC	subcutaneous
SCT	stem cell transplant
SD	stable disease
SMA	Safety Management Attachment to the Investigator's Brochure
$t_{1/2}$	terminal disposition half-life
TEAE	Treatment-emergent adverse event
TGI	tumor growth inhibition
T_{max}	single-dose time to reach maximum (peak) concentration
CCI	[redacted]
TTP	Time to (disease) progression
ULN	upper limit of the normal range
US	United States
VGPR	Very good partial response
WBC	white blood cell
WHO	World Health Organization

1. INTRODUCTION

In general, the purpose of the Statistical Analysis Plan (SAP) is to provide a framework that addresses the protocol objectives in a statistically rigorous fashion, with minimized bias or analytical deficiencies. Specifically, this plan has the following purpose:

To prospectively (a priori) outline the types of analyses and data presentations that will address the study objectives outlined in the protocol, and to explain in detail how the data will be handled and analyzed, adhering to commonly accepted standards and practices of biostatistical analysis in the pharmaceutical industry.

1.1 Study Design

This is a phase 3, randomized, double-blind, multicenter study to evaluate the safety and efficacy of MLN9708 versus placebo when added to lenalidomide and dexamethasone in patients with newly diagnosed multiple myeloma (NDMM). Patients must be previously untreated for symptomatic MM, be ineligible for high-dose therapy plus stem cell transplantation (HDT-SCT) because of age (ie, ≥ 65 years) or coexisting conditions per investigator judgment, be candidates for treatment with lenalidomide and dexamethasone as their standard therapy, and meet other eligibility criteria.

Following the Screening period, patients to be enrolled will be randomized to receive either MLN9708 or placebo in a double-blind fashion in addition to the background therapy of lenalidomide plus dexamethasone (LenDex). Eligible patients will be randomized in a 1:1 ratio into those 2 treatment arms, stratified by age (<75 years vs ≥75), ISS (stage I or II vs stage III), and BPI-SF worst pain score (<4 vs ≥4) at Screening.

Patients may continue to receive treatment for a maximum duration of 18 cycles (approximately 18 months with 28 days per cycle), or until progressive disease (PD) or unacceptable toxicity, whichever comes first. Patients remaining on study after 18 cycles will continue treatment in the same randomization arm on the same schedule with modified dose levels of the study drug and LenDex: reduce MLN9708 (or placebo) dose to 3.0 mg, reduce lenalidomide dose to 10 mg, and no dexamethasone.

Patients will be assessed for disease response and progression by an independent review committee (IRC). Response will be assessed according to the International Myeloma Working Group (IMWG) criteria for all patients every cycle during the treatment period and subsequently every 4 weeks during the PFS follow-up period until disease progression. All patients will be followed for survival after progression. Patients will be contacted every 12 weeks until death or termination of the study by the sponsor.

1.2 Study Objectives

The primary objective is:

- To determine whether the addition of oral MLN9708 to lenalidomide and dexamethasone improves progression-free survival (PFS) in patients with NDMM

The key secondary objectives are:

- To determine whether the addition of oral MLN9708 to lenalidomide and dexamethasone improves overall survival (OS)
- To determine whether the addition of oral MLN9708 to lenalidomide and dexamethasone improves the rate of complete response (CR)
- To determine whether the addition of oral MLN9708 to lenalidomide and dexamethasone improves pain response rate, as assessed by the Brief Pain Inventory – Short Form (BPI-SF) and analgesic use

Other secondary objectives are:

- To determine overall response rate (ORR), including partial response (PR), very good partial response (VGPR), and CR
- To determine time to response (TTR), duration of response (DOR), and time to progression (TTP)
- To determine the effect of the addition of MLN9708 to lenalidomide and dexamethasone on progression-free survival 2 (PFS2), defined as the date from randomization to the date of second disease progression or death from any cause, whichever comes first
- To determine the safety of the addition of MLN9708 to lenalidomide and dexamethasone
- To assess change in global health status, as measured by the global health status, functioning, and symptoms as measured by the patient-reported outcome (PRO) instrument European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) and MY20 module

- To determine the PFS and OS in high-risk cytogenetic patient groups defined by the following cytogenetic abnormalities: t(4;14), t(14;16), amp(1q21), and del(17p)

- To evaluate minimal residual disease status (MRD), via flow cytometry, in patients suspected to have reached CR at any time during the entire conduct of the study, and at Cycle 18 for patients who have maintained a CR until that point. The impact of MRD status on TTP, PFS, and OS will be assessed.

- To assess time to pain progression

- To collect pharmacokinetic (PK) data to contribute to population PK analyses

- To evaluate the frequency of skeletal-related events (eg, new fractures [including vertebral compression or rib fractures], irradiation of or surgery on bone, or spinal cord compression) from baseline through the last survival assessment

Exploratory Objectives are:

CCI

[redacted]

CCI

[redacted]

2. POPULATIONS FOR ANALYSIS

2.1 Intent-to-Treat Population

The Intent-to-Treat (ITT) population is defined as all patients who are randomized. Patients will be analyzed according to the treatment they were randomized to receive, regardless of any errors of dosing. If patients are regarded as screen failure, either were not randomized yet, or were randomized without being dosed, they will be excluded from ITT population.

The ITT population will be used for the primary, secondary efficacy analyses, and resource utilization and patient reported outcome analysis.

2.2 Safety Population

The safety population is defined as all patients who receive at least 1 dose of any study drug. Patients will be analyzed according to the treatment actually received. That is, those patients who are randomized to the active arm but receive the regimen in the control arm will be included in the control arm; those patients who are randomized to the control arm but receive the regimen in the active arm will be included in the active arm for safety analyses. More specifically, patients who received any dose of ixazomib will be included in the MLN9708 + LenDex arm and patients who did not receive any dose of ixazomib will be included in the placebo plus LenDex arm, regardless of their randomized treatment.

Safety population will be used for all safety related analyses such as AE, concomitant medication, laboratory tests, and vital signs.

2.3 Response-Evaluable Population:

The response-evaluable population was defined as all patients in the ITT population who receive at least 1 dose of any study drug, have measurable disease at baseline, and at least 1 post baseline response assessment assessed by an IRC. The response-evaluable population will be used for the analyses of time to response, and duration of response (defined in patients with confirmed response and will be summarized descriptively). Patients have measurable disease defined by at least 1 of the following 3 measurements:

- Serum M-protein ≥ 1 g/dL (≥ 10 g/L).
- Urine M-protein ≥ 200 mg/24 hours.
- Serum free light chain assay: involved free light chain level ≥ 10 mg/dL (≥ 100 mg/L) provided the serum free light chain ratio is abnormal.

2.4 Per-Protocol (PP) population

The PP population is a subset of the ITT population. The PP population consists of all patients who do not have major protocol violations, as determined by the study clinician, who is blinded to study drug assignment. All decisions to exclude patients from the PP population will be made before the unblinding of the study.

The PP population will be used as a sensitivity analysis of the ITT population for the primary efficacy endpoint PFS if there are more than 5% patients are excluded from the ITT population.

3. HYPOTHESES AND DECISION RULES

3.1 Statistical Hypotheses

There is one primary endpoint in this study. (See section 5.7.2 for study treatment arms)

The null and alternative hypothesis for PFS is:

H_0: PFS in Arm MLN9708+LenDex= PFS in Arm LenDex

H_a: PFS in Arm MLN9708+LenDex > PFS in Arm LenDex

There are three key secondary efficacy endpoints in this study.

The null and alternative hypothesis for OS is:

H_0: OS in Arm MLN9708+LenDex = OS in Arm LenDex

H_a: OS in Arm MLN9708+LenDex > OS in Arm LenDex

The null and alternative hypothesis for CR rate during the treatment period is:

H_0: CR rate in Arm MLN9708+LenDex = CR rate in Arm LenDex

H_a: CR rate in Arm MLN9708+LenDex > CR rate in Arm LenDex

The null and alternative hypothesis for pain response rate (analyzed in patients with baseline worst pain score ≥ 4) is:

H_0: Pain response rate in Arm MLN9708+LenDex = Pain response rate in Arm LenDex

H_a: Pain response rate in Arm MLN9708+LenDex > Pain response rate in Arm LenDex

3.2 Statistical Decision Rules

A closed sequential testing procedure will be used to test the primary endpoints and all 3 key secondary endpoints with the following testing order:

1. PFS (primary endpoint) in the ITT population at the first or both IAs and PFS at IA2 in 3 prespecified subgroups: 1) patients with baseline CrCl > 60 mL/min; 2) patients aged < 75 years; and 3) patients harboring expanded high-risk cytogenetic abnormalities defined as del(17p), t(4;14), t(14;16), and amp(1q21);
2. OS (first key secondary endpoint) at the IAs or FA;
3. CR rate (second key secondary endpoint) at the IAs or FA; and
4. Pain response rate (third key secondary endpoint) at the IAs or FA.

OS will be tested at the IAs or FA at the significance level determined by the O'Brien-Fleming alpha spending function (the Lan-DeMets method). The proof of strong control of the Type I error rate for testing PFS and OS in the ITT population and PFS in the subgroup populations is shown in the appendix in the SAP. CR rate will be tested at the same alpha level as that for OS whenever OS reaches statistical significance. Pain response rate will be

tested at the same alpha level as that for CR rate whenever CR rate reaches statistical significance. Due to the closed sequential testing property, the family-wise type I error is strongly controlled for both the primary endpoint and key secondary endpoints.

All other efficacy endpoints will be tested at a 2-sided alpha level of 0.05.

4. INTERIM ANALYSIS

4.1 Interim Analysis

There are 2 planned IAs. The first IA will be performed when approximately 326 disease progression/death events have occurred. This IA is expected to occur approximately 45 months after the first patient is enrolled. If the test for PFS in the ITT population is statistically significant at the first IA, this will be the FA for PFS for statistical testing purposes, central efficacy and investigator assessments of disease response for protocol purposes will be discontinued (except for investigator assessment of PFS2) given that the primary endpoint has been met, and the second IA will be conducted for OS when approximately 250 death events have occurred. If the test for PFS does not reach statistical significance at IA1 in the ITT population, then at IA2 PFS will be tested in both the ITT population and in 3 prespecified subgroups, as described below.

The subgroup testing strategy approach includes 2 major components: a) preservation of the ability to detect the overall treatment effect using a reduced overall significance level of $\alpha_1 = 0.04$, which will be used for the ITT population, and b) test of treatment effect for the 3 prespecified subgroups: 1) patients with baseline CrCl > 60 mL/min; 2) patients aged < 75 years; and 3) patients harboring expanded high-risk cytogenetic abnormalities defined as del(17p), t(4;14), t(14;16), and amp(1q21). Subgroup testing will be conducted using the remaining $\alpha_2 = 0.01$ and the Hochberg procedure for multiplicity correction among the 3 prespecified subgroups (refer to the appendix in the SAP for proof of strong control of the Type I error rate). Because the size of the treatment effect may be substantially greater in a prespecified subgroup than in the overall study population, analysis of patients in each subgroup at a stringent significance level may still provide a statistically significant outcome. The detailed statistical design schema is presented in Figure 4-1.

Figure 4-1 Schematic of Statistical Plan

For the testing of PFS in the ITT population, the Gamma(-1) alpha spending function will be used to calculate the significance boundary based on the observed number of PFS events with total alpha=0.04. The first IA will be performed when approximately 326 PFS events have occurred. This will be the first analysis for PFS for statistical testing purposes. If the test is statistically significant, then this analysis will be the FA of PFS for statistical testing purposes. No subsequent PFS testing will be conducted, and central efficacy and investigator assessments of disease response for protocol purposes will be discontinued except for the investigator assessment of PFS2 (see the Schedule of Events of protocol). In this scenario, the second IA will be for OS testing when approximately 250 death events have occurred and will determine whether the final number of OS events might be increased.

If the test for ITT PFS is not statistically significant at the first IA, response assessments will continue until IA2, and PFS testing in the ITT and subgroup populations will be conducted in parallel at the second IA, when approximately 370 PFS events have occurred (rather than the previous study design of 435 PFS events); this will be the FA of PFS for statistical testing purposes. If the test for PFS is significant at the second IA, OS will be tested, and determination of whether the final number of OS events will be increased from 320 to up to

400 will occur. If the test for PFS in the second IA is not statistically significant in any population (the ITT or any of the 3 subgroups), the study may be stopped.

Because at the time of this amendment, the boundary for ITT PFS at IA1 has already been calculated based on 328 PFS events observed at IA1, 435 PFS events targeted at PFS final analysis, and the Gamma(-1) alpha-spending function, this boundary will not be changed. However, the boundary for ITT PFS at IA2 (final analysis of ITT PFS) will be calculated based on the observed number of PFS events at IA2 in order to spend what is left of the overall alpha-level 0.04 for ITT. The final boundaries at IA1 and IA2 will not approximate a Gamma(-1) function, but type I error will remain protected under the flexible alpha-spending approach (see appendix in the SAP for more details).

For the testing of OS, alpha spending for IA1 and IA2 will always be based on the observed events (information fraction) using alpha=0.04 with a different adjustment of critical value at OS FA testing (CHW test statistics [3] will be used for the primary analysis of OS at FA) based on the following scenarios:

1. If ITT PFS is significant in IA1, then ITT OS will be tested in the FA with a total alpha of 0.04; there is no test on subgroup PFS.
2. If ITT PFS is not significant in IA1, then parallel testing of the ITT population PFS and the subgroup populations PFS will occur in IA2:
    a. If the ITT population's PFS is significant and at least 1 subgroup is not significant, then the ITT population's OS will be tested at IA2 and FA with potential sample size re-estimation using a total alpha of 0.04.
    b. If the ITT population's PFS is significant and all 3 subgroup populations' PFS are significant, then the ITT population's OS will be tested at IA2 and FA where the critical value at FA can be updated based on a total alpha of 0.05.
    c. If the ITT population's PFS is not significant and at least 1 subgroup population's PFS is significant, then no formal ITT OS testing will be conducted.

The family-wise error rate for the 4 null hypotheses for PFS and the 1 hypothesis for OS for the overall study population is controlled using a prespecified, 2-sided 0.05 level of significance. The proof of strong control of the Type I error rate for testing PFS and OS in the ITT population and PFS in the subgroup populations is shown in the appendix in the SAP. For the other 2 key secondary endpoints, the CR rate will be tested at the same alpha

level, instead of the same critical value, as that of the OS analysis when OS reaches statistical significance. The pain response rate will be tested at the same alpha level as that of the CR rate analysis when the CR rate reaches statistical significance. Because of the closed sequential testing property, the family-wise error rate is strongly controlled for both the primary endpoint and the 3 key secondary endpoints [4].

The IAs will be conducted by the independent statistical center (ISC) and presented for review to the IDMC. During the closed session of the IDMC meeting at IA2, the IDMC will compare the conditional power for OS based on the interim result with the prespecified adaptation rules and recommend to the sponsor executive committee the final adaptation decision on OS. The adaptation rule will be included in the appendix of IDMC charter and can only be accessed by ISC, IDMC, Head of Biostatistics and the sponsor design statistician who are not involved in the study conduct. This recommendation will be documented in the IDMC closed meeting minutes.

4.2 Independent Data Monitoring Committee (IDMC)

An IDMC supported by an independent statistician will review safety at regular intervals additionally safety and efficacy data at 2 planned interim analyses. The IDMC will provide a recommendation regarding study continuation based on the safety and efficacy parameters. In the event that the study is terminated early based on the IDMC recommendation, Millennium will notify the appropriate regulatory authorities. In addition, the IDMC will periodically review safety data at regularly scheduled meetings prespecified in the IDMC charter.

The first formal safety review will occur after approximately 60 subjects have been randomized and receive at least 1 cycle of study treatment. Subsequently, periodic safety reviews will also occur as prespecified in the IDMC charter.

Study accrual will not be interrupted due to the scheduled safety reviews. The IDMC or MLN9708 study team may request an ad hoc meeting for any reason, including a significant unexpected safety event, unplanned unblinding of study results, follow-up of an observation during a planned IDMC meeting, or a report external to the study, such as publication of study results from a competing product. At each review, subject incidence rates of AEs (including all serious AEs, treatment-related AEs, serious treatment-related events, and events requiring the discontinuation of study drug) will be tabulated by System Organ Class, preferred term, and severity grade. Listings and/or narratives of "on-study" deaths and other serious and significant AEs, including any early withdrawals due to AEs, will be provided.

Records of all meetings will be archived. The IDMC will communicate major safety concerns and recommendations regarding study modification or termination to Millennium. Further details will be provided in the IDMC charter.

4.3 Independent Review Committee (IRC)

An independent review committee (IRC) will review all blinded disease evaluation data from the study and determine disease status (response and progression). Data from the IRC will not be provided back to the investigator during the conduct of the study. Likewise, investigator response assessments will not be provided to the IRC.

5. STATISTICAL METHODOLOGY

In general, summary tabulations will be presented by treatment arm and will display the number of observations, mean, standard deviation, median, minimum, and maximum for continuous variables, and the number and percent per category for categorical data. The Kaplan-Meier survival curves and 25th, 50th (median), and 75th percentiles will be provided along with their 2-sided 95% CIs for time-to-event data.

5.1 Sample Size Justification

The primary objective of this study is to determine if MLN9708 plus lenalidomide and dexamethasone improves PFS compared with placebo plus lenalidomide and dexamethasone in patients with newly diagnosed MM. The study will not be stopped after the PFS analysis, however, even if a significant PFS is observed, in order to obtain an adequate statistical power for OS.

The total sample size was calculated based on maintaining 80% power to test the OS. The study is also adequately powered to test PFS. There is 2 planned IA and 1 FA.

Assuming a hazard ratio of 0.70 (median PFS of 25 months in control arm versus 35.8 months in treatment arm), 370 PFS events will be needed (92% power and 2-sided alpha of 0.04) with up to 2 planned PFS analyses conducted as described in the section 4.1.

The first IA will be performed when approximately 326 PFS events have occurred. This is expected to occur approximately 45 months after the first patient is enrolled, including a 27-month enrollment period and additional 18-month follow-up from the last patient.

If the test for PFS in the ITT population is statistically significant at the first IA, this will be the FA for PFS for statistical testing purposes, and the second IA will assess OS when approximately 250 death events have occurred.

If the test for PFS in the ITT is not statistically significant at the first IA, then the second IA will assess PFS and OS when approximately 370 PFS events have occurred. In addition, in such a case, PFS will be tested at IA2 in 3 prespecified subgroups: 1) patients with baseline CrCl > 60 mL/min; 2) patients aged < 75 years; and 3) patients harboring expanded high-risk cytogenetic abnormalities defined as del(17p), t(4;14), t(14;16), and amp(1q21).

For the final OS analysis, the total event size calculation will be based on the adaptive sample size re-assessment approach.[3, 5] The minimum event size of 320 death events is based on an optimistic assumption of a hazard ratio of 0.72 (median survival of 50 months in the control arm vs 69.4 months in the treatment arm) with 80% power at a 2-sided 0.05 level of significance. The O'Brien-Fleming alpha spending function (the Lan-DeMets method) will be used to calculate the significance boundary based on observed number of death events in each IA with a total of 320 OS events for the FA. In the second IA, if OS significance is not claimed, the conditional power based on OS will be calculated. If the conditional power falls in the favorable zone or unfavorable zone, the FA of OS with approximately 320 events will remain unchanged. If the conditional power falls in the promising zone, the event size will be determined according to a prespecified sample size adaptation rule, with an event cap of 400 OS events. No futility analysis will be performed in the study.

5.2 Randomization and Stratification

Randomization scheme will be generated by an independent statistician at Millennium who is not on the study team. Prior to dosing, a randomization number will be assigned to each patient. The randomization assignment will be implemented by an interactive voice/ web response system (IXRS).

Eligible patients will be randomized in a 1:1 ratio into those 2 treatment arms, stratified by: age (<75 years vs ≥ 75), ISS (stage 1 or 2 vs stage 3), and BPI-SF worst pain score (< 4 vs ≥ 4) at screening.

5.3 Blinding and Unblinding

This is a double-blind study: all study personnel including the investigators, site personnel, study clinicians, and the sponsor will be blinded to the treatment assignments for the duration of the study. Only the independent statistical center (ISC) and IDMC will have access to un-blinded individual patient level data. The periodic safety analyses will be generated for the IDMC by an ISC. The formal interim efficacy analyses will also be conducted by ISC for the IDMC.

Refer to section 4.2 for the roles and responsibilities of IDMC.

5.4 Data Handling

5.4.1 Methods for Handling Missing Data

All available efficacy and safety data will be included in data listings and tabulations. Data that are potentially spurious or erroneous will be examined according to standard data management operating procedures.

In general, missing data will be treated as missing and no data imputation will be applied, unless otherwise specified. For patient reported outcomes data, primarily missing data imputation will be based on published instrument specific methods. Other missing data imputation method such as Last Observation Carry Forward (LOCF) and multiple imputation method may be explored as sensitivity analyses for patient reported outcomes data.

For the key secondary endpoints CR rate, missing value is defined as no post-baseline response assessment either due to lost to follow-up or withdrawal by patient. In the primary analysis, if the response assessment in either arm is missing on comparing response rates, it will be counted as a failure (non-responder) instead of a missing value. The procedure to deal with missing data in the primary analysis for the pain response rate will be using the same method as CR rate.

5.4.1.1 Missing/Partial Dates in Screening Visit

The following rules apply to dates recorded in the screening visits.

- If only the day-component is missing, the first day of the month will be used if the year and the month are the same as those for the first dose of study drug. Otherwise, the 15th will be used.

- If only a year is present, and it is the same as the year of the first dose of study drug, the 15th of January will be used unless it is later than the first dose, in which case the date of the first of January will be used, unless other data indicates that the date is earlier.

- If only a year is present, and it is not the same as the year of the first dose of study drug, the 15th of June will be used, unless other data indicates that the date is earlier.

5.4.1.2 Missing/Partial Dates in Adverse Events/Concomitant Therapies/Subsequent Therapies

Every effort will be made to avoid missing/partial dates in on-study data.

Adverse events with stop dates that are completely or partially missing will be imputed as follows:

- If the stop date has month and year but day is missing, the last day of the month will be imputed
- If the stop date has year, but day and month are missing, the 31th of December will be imputed

After the imputation, the imputed dates will be compared against the date of death, if available. If the date is later than the date of death, the date of death will be used as the imputed date instead.

Adverse events with start dates that are completely or partially missing will be imputed as follows:

- If the start date has month and year but day is missing, the first day of the month will be imputed
    - If this date is earlier than the first dose date, then the first dose date will be used instead

  - If this date is later than the stop date (possibly imputed), then the stop date will be used instead

- If the start date has year, but day and month are missing, the 15th of June will be imputed
  - If this date is earlier than the first dose date, then the first dose date will be used instead
  - If this date is later than the stop date (possibly imputed), then the stop date will be used instead

If the start date of an event is completely missing, then it is imputed with the first dose date.

Concomitant therapies with start dates that are completely or partially missing will be analyzed as follows:

- If the start date has month and year but day is missing, the therapy will be included in the summary table if the month and year of the start date of the event are:
  - On or after the month and year of the date of the first dose of study drug

  and

  - On or before the month and year of the date of the last dose of study drug plus 30 days.
- If the start date has year, but day and month are missing, the therapy will be included in the summary table if the year of the start date of the event is:
  - On or after the year of the date of the first dose of study drug

  and

  - On or before the year of the date of the last dose of study drug plus 30 days.

If the start date of an event is completely missing, then the therapy will be included in the summary table.

Subsequent therapies with start dates that are completely or partially missing will be analyzed as follows:

- When month and year are present and the day of the month is missing,
  - If the onset month and year are the same as the month and year of last dose with study drug, the day of last dose + 1 will be imputed.

  - If the onset month and year are not the same as the month and year of last dose with study drug, the first day of the month is imputed.

- When only a year is present,

  - If the onset year is the same as the year of last dose with study drug, the date of last dose + 1 will be imputed.

  - If the onset year is not the same as the year of last dose with study drug, the first day of the year is imputed.

- If no components of the onset date are present the date of last dose + 1 will be imputed.

5.4.2 Definition of Baseline Values

Unless otherwise specified, the baseline value is defined as the value collected at the time closest to, but prior to, the start of study drug administration.

5.4.3 Windowing of Visits

All data will be categorized based on the scheduled visit at which it was collected. These visit designators are predefined values that appear as part of the visit tab in the eCRF.

5.4.4 Pooling

All data from all sites will be pooled. Study center or treatment-by-center interaction will not be included in any statistical analysis.

5.4.5 Withdrawals, Dropouts, Loss to Follow-up

Time to event parameters will be censored if patients withdraw, drop out, or are lost to follow-up before documentation of the events (progressive disease / death). Rules for censoring are detailed in section 5.8.

5.5 Patient Disposition

Patient disposition includes the number and percentage of patients for the following categories: patients in each of the study populations, patients discontinued from the treatment, primary reason to discontinue from the treatment, patients discontinued from the study, and primary reason to discontinue from the study. All percentages will be based on the number of patients in the ITT population.

A listing will present data concerning patient disposition.

5.6 Demographics and Baseline Disease Characteristics

5.6.1 Demographics

Demographics will be summarized by treatment groups in a descriptive fashion in the ITT population. Baseline demographic data to be evaluated will include age, sex, race, ethnicity, height, weight, and other parameters as appropriate. Patient enrollment by region and country will also be summarized by treatment groups.

5.6.2 Medical History

General medical history and prior medications will be listed for all patients.

Medical history will be summarized (frequency and percentage) for both treatment groups by the disease categories recorded in the database. A patient is counted only once within a category. Percentages are based on the number of patients in the safety population within each treatment group.

5.6.3 Baseline Disease Status

Baseline disease characteristics (Eastern Cooperative Oncology Group [ECOG]) performance status, co-morbidity status by age (<65, 65≤ age <75, ≥75), type of myeloma, ISS stage, serum M-protein, urine M-protein, β_2-microglobin by category (ie, < 2.5, 2.5-5.5, > 5.5 mg/L), serum creatinine and its category (≤ 2, >2 mg/dL), creatinine clearance by category (ie, >30-60, >60 mL/min), serum albumin by category (ie, < 3.5, ≥ 3.5 g/dL), corrected calcium, Durie-Salmon stage, Lytic bone lesions, extramedullary disease will be summarized for all patients. Months from initial diagnosis to first dose of MLN9708 will be summarized for all patients if there is sufficient data for analysis.

A patient's type of myeloma is determined by the combination of heavy chain type (IgG, IgA, IgM, IgD, IgE, and other) and light chain type (Kappa, Lambda, and biclonal). In descriptive summaries, Myeloma type will be summarized separately for the heavy chain patients (according to IgG, IgA, IgM, IgD, IgE, biclonal, other) and for the light chain patients (according to kappa or lambda or biclonal).

Creatinine clearance is to be calculated using the Cockcroft-Gault formulas as follows:

For male patients:

$$\text{creatinine clearance} = \frac{(140 - \text{Age[yrs]}) \times \text{weight[kg]}}{72 \times (\text{serum creatinine [mg/dL]})}$$

For female patients:

$$\text{creatinine clearance} = 0.85 \times \frac{(140 - \text{Age[yrs]}) \times \text{weight[kg]}}{72 \times (\text{serum creatinine [mg/dL]})}$$

Integer values will be used.

Months from diagnosis to the randomization date for each treatment is calculated by

$$\frac{\text{randomization date - date of diagnosis}}{365.25/12}$$

Distribution of stratification factors will also be summarized.

5.6.3.1 Extent of disease at baseline

The following categories of extent of disease at baseline will be summarized: number of patients with bone marrow aspirate, bone marrow aspirate results (% plasma cells, % megakaryocytes present), number of patients with bone marrow biopsy, bone marrow biopsy results (% plasma cells, % cellularity, type of cellularity, % Kappa/Lambda ratio performed), skeletal survey results and imaging including Magnetic Resonance Imaging/ Computed Tomography/PET-CT results (normal, abnormal not clinically significant, abnormal clinically significant, and not done), number and percentage of present lytic bone lesions, number of extramedullary plasmacytoma, type of extramedullary plasmacytoma.

Percentage for all categorical summarizations for bone marrow biopsy/aspirate and aspirate is based on patients with an adequate sample for the specified test.

5.6.4 Bone Marrow Cytogenetic Results at Baseline

Bone marrow cytogenetic results at baseline from the conventional/karyotype and molecular/FISH cytogenetic analyses methods will be displayed. The results will be categorized as "Normal", "Abnormal" and "Indeterminate". The percentage of each category will be summarized.

The following are the categories of interest:

1. Del 17 positive group (made up of del 17 alone or in combination with t(4;14) or t(14;16) or amp(1q21))

2. t(4;14) alone [no del 17, t(4;14), t(14;16) or amp(1q21)]

3. t(14;16) alone [no del17, t(4;14), t(14;16) or amp(1q21)]

4. amp(1q21) alone [no del17, t(4;14) or t(14;16)]

5. High risk group: made up of del17, t(4;14) or t(14;16)

6. Expanded High risk group: made up of del17, t(4:14), t(14;16) or amp(1q21)

Standard risk group definition will differ for the high risk and the expanded high risk group and will be defined as patients for whom the tests for del17, t(4;14), t(14;16) and amp(1q21) are normal. Detailed definitions are listed in the section 5.8.1.1 on definition of subgroup.

Abnormal types of interest, including but not limited to del 13, del 17, t(4;14), t(14;16), will also be tabulated.

5.7 Treatments and Medications

5.7.1 Concomitant Medications

Concomitant medications will be coded by preferred term using the World Health Organization (WHO) Drug Dictionary. The number and percentage of patients taking concomitant medications from the first dose through the end of the on-treatment period will be tabulated by Anatomical Therapeutic Chemical (ATC) classification pharmacological subgroup and WHO drug preferred term for each treatment group in the safety population. By-patient listing will also be presented for concomitant medications.

Concomitant procedures will not be coded, but will be presented in a data listing in the safety population.

Types of subsequent therapy will also be summarized accordingly in the table and listing.

5.7.2 Study Treatments

Following the Screening period, patients who will be enrolled and treated with lenalidomide plus dexamethasone will be randomized to receive a study drug in a double-blind fashion, either MLN9708 or placebo. Eligible patients will be randomized in a 1:1 ratio into those 2 treatment arms.

Arm MLN9708+LenDex: Patients will receive MLN9708 4.0 mg capsule on Days 1, 8, and 15 plus lenalidomide (25 mg) on Days 1 through 21 and dexamethasone (40 mg) on Days 1, 8, 15, and 22 of a 28-day cycle.

Arm LenDex: Patients will receive placebo capsule on Days 1, 8, and 15 plus lenalidomide (25 mg) on Days 1 through 21 and dexamethasone (40 mg) on Days 1, 8, 15, and 22 of a 28-day cycle.

In both arms, patients over 75 years of age will receive reduced dexamethasone dose (20mg). Dose modifications may be made throughout the study based on toxicities. Patients with a low creatinine clearance ≤ 60 mL/min (or ≤ 50 mL/min, according to local label/practice) will receive a reduced lenalidomide dose of 10 mg once daily on Days 1 through 21 of a 28-day cycle. The lenalidomide dose may be escalated to 15 mg once daily after 2 cycles if the patient is not responding to treatment and is tolerating the treatment. If renal function normalizes (ie, creatinine clearance > 60 mL/min or > 50 mL/min, according to local label/practice) and the patient continues to tolerate this treatment, lenalidomide may then be escalated to 25 mg once daily.

Patients may continue to receive treatment as outlined previously for 18 cycles (approximately 18 months), or until progressive disease (PD) or unacceptable toxicity, whichever comes first. After 18 cycles, patients will continue treatment in the same randomization arm on the same schedule with modified dose levels of the study drug and LenDex: reduce MLN9708 (or placebo) dose to 3.0 mg, reduce lenalidomide dose to 10 mg, and no dexamethasone.

5.7.2.1 Duration of Follow-up

The duration of follow-up is defined as time from randomization to the death or last known visit. If a subject dies, the duration equal to date of death minus study start + 1 with censor variable =1 (censored for follow up). If a subject is alive, the duration equal to the date subject last known to be alive minus study start + 1 with censor variable=0 (event for follow up).

Duration of follow-up for maintenance portion is defined as time from the date of first dose of maintenance to the death or last known visit.

5.7.2.2 Extent of Exposure

An overall summary of drug exposure will be presented including number of treated cycles, numbers and percentages of patients who had ≥1, ≥2, …, and ≥36 treated cycles, for each treatment group in the safety population. Aggregate summary of numbers and percentages of patients who had 1-6, 7-12, 13-18, 19-24, 25-30, 31-36, ≥37 treated cycles will also be

presented in the same table. Extent of Exposure (days), which is calculated as (Last Dose Date of study drug – First Dose Date of study drug + 1), will also be presented.

Additionally exposure to dexamethasone will be characterized by total amount of dose taken in mg, total number of dose taken, number of treated cycles, numbers and percentages of patients who had ≥1, ≥2, …, and ≥36 treated cycles, and relative dose intensity (%) for each treatment group in the safety population. Aggregate summary of numbers and percentages of patients who had 1-6, 7-12, 13-18, 19-24, 25-30, 31-36, ≥37 treated cycles will also be presented in the same table.

MLN9708 and lenalidomide exposure will be summarized similarly as dexamethasone for the applicable treatment group/option.

A treated cycle is defined as a cycle in which the patient received any amount of any study drug.

A treated cycle for a specific drug is defined as a cycle in which the patient received any amount of the specific drug.

Relative dose intensity (RDI) (%) is defined as 100 x (total dose received in mg) / (sum of prescribed dose over all treated cycles). For prescribed dose, if patients with a low creatinine clearance ≤ 60 mL/min received reduced lenalidomide dose of 10 mg at C1D1, then 10 mg will be used in the denominator per protocol dosing administration. Similarly, 20 mg will be used for Dexamethasone RDI calculation for patients over 75 years old. After 18 cycles, MLN9708 will be reduced to3 mg, Len will be reduced to 10 mg daily and Dex will be discontinued, so prescribed dose per protocol will be updated and reflected in the calculation accordingly.

Dosing data will also be presented in a by-patient listing.

5.7.2.3 Treatment Modifications

Action on each study drug will be summarized by each of the Cycle 1 through 36, sum of the remainder Cycles, Cycles 1-6, Cycles 7-12, Cycle 13-18, Cycles 19-24, Cycle 25-30, Cycle 31-36, >=37 and total for each treatment group in the safety population.

5.8 Efficacy Analyses

All efficacy evaluations will be conducted using the ITT population unless otherwise specified.

5.8.1 Primary Efficacy Endpoint

There is 1 primary endpoint: PFS, which is defined as the time from the date of randomization to the date of first documentation of PD or death due to any cause, whichever occurs first. Patients without documentation of PD will be censored at the date of last response assessment. The details regarding the handling of missing assessment and censoring for PFS analysis are presented in Table 5-1.

Table 5-1 Handling of Missing Assessment and Censoring for PFS Primary Analysis based on FDA guidance

Situation	Date of Progression or Censoring	Outcome
No baseline and/or no post baseline assessment, no subsequent anticancer therapy after study treatment, no death	Date of Randomization	Censored
Disease progression documented between scheduled visits	Date of documented disease progression	Event
No documented death or disease progression	Date of last adequate assessment*	Censored
Lost to follow-up, withdraw consent before any documented death or disease progression	Date of last adequate assessment*	Censored
Death or progression after more than one missed visit	Date of last adequate assessment*	Censored
Alternate antineoplastic therapy started prior to disease progression	Date of last adequate assessment prior to starting alternate antineoplastic therapy	Censored
Death before first assessment	Date of death	Event
Death between adequate assessment visits	Date of death	Event

* Adequate disease assessment is defined as there is sufficient data to evaluate a patient's disease status.

5.8.1.1 Primary Efficacy Analysis

PFS will be analyzed when approximately 326 PFS events have occurred. A 2-sided, stratified log-rank test will be used to compare the treatment groups with respect to PFS at a 2-sided alpha level of 0.05. In addition, an unadjusted stratified Cox model will be used to estimate the hazard ratio and its 95% CIs for the treatment effect using the stratification factors. The Kaplan Meier (K-M) survival curves and K-M medians (if estimable), along with their 2-sided 95% CIs, will also be provided for each treatment group.

Sensitivity analyses for PFS include:

1. PFS assessed by investigator will be analyzed in the ITT population.

2. PFS assessed by IRC will be analyzed in the per protocol population.

PFS assessed by IRC using different censoring mechanisms will be analyzed in the ITT population, for example, not censoring for patients who discontinue treatment and go on transplant or alternative antineoplastic therapy. The other details of the handling of missing assessment and censoring for additional sensitivity analyses are presented in Table 5-2. Sensitivity analyses will be performed on the basis of one alteration at a time, not on combined alterations unless specified otherwise. Additional sensitivity analysis for PFS might be conducted on treating start date of alternate antineoplastic therapy as events.

Table 5-2 Handling of missing assessment and censoring for PFS Sensitivity Analysis based on EMA guidance

Situation	Date of Progression or Censoring	Outcome
Alternate antineoplastic therapy started prior to disease progression	Date of documented disease progression	Event
Death or disease progression after more than one missed visit	Date of death or disease progression	Event

Subgroup analyses will be performed for PFS relative to baseline stratification factors, demographic data such as sex, race, region (e.g. North America, Europe and Other), and disease characteristics, CCI [redacted] The details on subgroups are presented in the following:

Subgroup	Definition of Group
Age	< 75 years, ≥ 75 years
Sex	male vs female
Race	white, black-African American, Asian, other
Region	North America, Europe, APAC, other
Cytogenetic risk	Standard-risk[1], high-risk [(del17); t(4;14); t(14;16)], not available Standard-risk[2], expanded high-risk [(del17); t(4;14); t(14;16); amp(1q21)], not available
ISS stage	In additional to stratification factors , also define as I or II or III
Renal function based on baseline creatinine clearance	< 60 mL/min, and ≥60 mL/min
ECOG performance status	0 or 1 vs 2
CCI [redacted]	

1. Standard Risk in this analysis is defined as del (17), t(4:14) and t(14:16) normal
2. Standard Risk in this analysis is defined as del (17), t(4:14) , t(14:16) and 1q 21 normal

5.8.2 Key Secondary Efficacy Endpoints

There are 3 key secondary endpoints: CR rate, OS and Pain response rate.

Overall Survival

OS is defined as the time from the date of randomization to the date of death. Patients without documentation of death at the time of analysis will be censored at the date last known to be alive. OS will be analyzed based on the ITT population.

CR Rate

The CR rate is defined as the proportion of patients who achieve CR assessed by an IRC relative to the ITT population during the treatment period. If the response assessment in either arm is missing on comparing CR rates, it will be counted as a failure (non-responder) instead of a missing value.

Pain Response Rate

Pain response is defined, among patients whose baseline pain score are >=4, as the occurrence of at least a 30% reduction from baseline in BPI-SF worst pain score over the last 24 hours without an increase in analgesic use for 2 consecutive measurements >=28 days apart.

5.8.2.1 Key Secondary Efficacy Analysis

Three key secondary efficacy endpoints will be tested sequentially in the order of 1) OS; 2) CR rate; 3) Pain response rate. OS will be tested at the IAs or FA at the significance level determined by the O'Brien-Fleming alpha spending function (the Lan-DeMets method). CR rate will be tested at the same alpha level as that for OS whenever OS reaches statistical significance. Pain response rate will be tested at the same alpha level as that for CR rate whenever CR rate reaches statistical significance. Due to the closed sequential testing property, the family-wise type I error is strongly controlled for both the primary endpoint and key secondary endpoints.

Overall Survival

A 2-sided, stratified log-rank test will be used to compare the treatment groups with respect to OS. The test significance level at the IA and FA is decided by the O'Brien-Fleming alpha spending function (the Lan-DeMets method). In addition, an unadjusted stratified Cox

model will be used to estimate the hazard ratio and its 95% CIs for the treatment effect using the stratification factors. The K-M survival curves and K-M medians (if estimable), along with their 2-sided 95% CIs, will also be provided for each treatment group.

To adjust for the potential confounding effects of subsequent therapies after patients discontinue study treatment, the following 2 methods will be used:

- Marginal Structural Models (MSMs) by Robins and Finkelstein [2000]
- Inverse Probability of Censoring Weighted (IPCW) method by Robins and Finkelstein [2000]

In the MSM and IPCW analyses, in order to derive weights adjusting for the time-fixed and time-varying confounding effects due to taking alternative therapies, the covariates that affect disease progression and post-progression treatment, and the OS endpoint will be used. Baseline covariates include region (North America, others), age (< 75, ≥ 75), race (white, non-white), ECOG score (0 or 1, 2), type of myeloma (IgA, other), presence of extramedullary plasmacytomas (yes, no), presence of lytic bone lesions (yes, no), cytogenetic abnormalities (high risk, others), baseline hemoglobin, baseline platelets, baseline creatinine clearance, baseline albumin, baseline LDH, baseline β_2 microglobulin, and baseline corrected calcium. Time-varying covariates include duration of exposure, disease progression status at each study visit, hemoglobin value at each study visit and progression/relapse, platelets value at each study visit and progression/relapse, M-protein value at each study visit and progression/relapse, and MRD status over time. The final criteria for selected covariates would need to be statistically have a p-value of less than or equal to 0.1 in the multivariate logistic regression models for weight calculations. If there are more than 5% missing in the baseline covariate, then this covariate will be dropped from the weighting calculation and final OS model. For both MSM and IPCW analyses, logistic regression models on repeated measurements will be used to approximate the Cox models in the weight derivations from which stabilized weights will be derived per subject per observation. SAS proc PHREG procedure with counting process type of data input, which takes multiple observations per subject, will be used as the final Cox model for OS for both MSM and IPCW approaches, where robust variance will be used to accommodate covariance introduced by correlated longitudinal observations within each subjects and other extra variabilities due to departure from model assumptions. Adjusted HRs, their

corresponding 95% confidence intervals, and adjusted p-values will be presented. Specific to MSM, interaction between active treatment and alternative therapy will be included in the final model if the p-value for this term is <0.1. For IPCW, adjusted K-M curves will be presented.

Subgroup analyses will be performed for OS, similarly as detailed in section 5.8.1.1 of PFS analysis.

CR Rate

CR rate will only be tested after statistical significance is achieved for PFS and OS. Stratified Cochran-Mantel-Haenszel (CMH) test will be used to compare CR rates between the 2 treatment arms. A logistic regression model will be used to estimate the treatment effect in terms of odds ratio. The odds ratio and its associated 95% CIs will be presented.

Sensitivity analyses for CR rate include but are not limited to:

1. Response assessed by investigator in the ITT population

2. Response assessed by IRC in the per protocol population

3. Response assessed by IRC in the response evaluable population

Pain Response Rate

If CR is significant, then Pain response rate will be analyzed in patients with baseline worst pain score ≥ 4 in the ITT population. Pain response rate is the proportion of patients who have a pain response and will be summarized by treatment groups. If the pain assessment in either arm is missing on comparing pain response rates, it will be counted as a failure (non-responder) instead of a missing value. The stratified CMH test will be used to compare the 2 treatment arms. In addition, the absolute treatment difference in pain response rate will be provided, along with 95% CI.

CCI

5.8.3 Other Secondary Efficacy Endpoints and Analyses

Other secondary efficacy parameters include overall response rate (ORR) , time to response (TTR), time to progression, duration of response, OS and PFS in high-risk population defined by del(17), and translocation t(4;14) and t(14;16) (at least one of these abnormalities), and expanded high-risk population defined as del(17), amp(1q21), and translocation t(4;14) and t(14;16) (at least one of these abnormalities)

Disease response-related endpoints will be analyzed using IRC-assessed response rate.

ORR

ORR is defined as the proportion of patients who achieved PR or better relative to the ITT population. ORR will be analyzed based on the ITT population using the method similar to that used in the CR rate analysis. Additional analysis will also be presented for CR+VGPR.

Time to Response

Time to response is defined as the time from randomization to the first documentation of PR or better. Time to response will be compared in the ITT population and summarized descriptively for the responders.

Time to Progression

TTP is defined as the time from the date of randomization to the date of first documentation of PD. Patients without documentation of PD at the time of analysis will be censored at the date of last response assessment that is SD or better. TTP will be analyzed based on the ITT population using the similar method as PFS.

Duration of Response

DOR is defined as the time from the date of first documentation of a PR or better to the date of first documentation of PD for responders. Responders without documentation of PD will be censored at the date of last response assessment that is SD or better. DOR will be summarized descriptively using the Kaplan-Meier method.

Progression-free survival 2

Progression-free survival 2 (PFS2) is defined as the time from the date of randomization to the date of first documentation of PD on the next antineoplastic therapy following study treatment or death due to any cause, whichever occurs first.

PFS2 will be analyzed based on the ITT population. A 2-sided, stratified log-rank test will be used to compare the treatment groups with respect to PFS2 at a 2-sided alpha level of 0.05. In addition, an unadjusted stratified Cox model will be used to estimate the hazard ratio and its 95% CIs for the treatment effect using the stratification factors. The Kaplan Meier (K-M) survival curves and K-M medians (if estimable), along with their 2-sided 95% CIs, will also be provided for each treatment group.

The details of the handling of missing assessment and censoring are presented in Table 5-3 and Table 5-4.

Table 5-3 Censoring for PFS2 For Those Who have Received Second line Therapy following Study Treatment

Situation	Date of Progression or Censoring	Outcome
Documented death or disease progression during second line therapy	Date of death/disease progression	Event
No documented death or disease progression during second line therapy	Date of last disease assessment	Censored
Lost to follow-up, withdraw consent before any documented death or disease progression during second line therapy	Date of last disease assessment	Censored
Start of third line therapy prior to the disease progression during second line therapy	Date of last disease assessment prior to starting the third line therapy	Censored

Table 5-4 Censoring for PFS2 for Those Who have not received Second Line of Therapy

Situation	Date of Progression or Censoring	Outcome
No documented death	Date of last visit	Censored
Death	Date of death	Event

Clinical Outcomes in High-Risk Population

Overall survival, PFS, ORR and DOR in the high-risk subgroups will be analyzed using a similar method as those in the ITT population. The following high-risk populations will be analyzed:

- By individual abnormality group within high risk: patients carrying 1 of the following cytogenetic abnormalities: del(17), translocation t(4;14), t(14;16)

  - the del17 will include pts with del17 alone along with pts where the del17 is associated to t(4;14),or t(14;16)
  - the t(4;14) group will include ONLY pts with t(4;14) ALONE (no del17, t(14;16))
  - the t(14;16) group will include ONLY pts with t(14;16) ALONE (no del17, t(4;14))

- By individual abnormality group within expanded high risk: patients carrying 1 of the following cytogenetic abnormalities: del(17), translocation t(4;14), t(14;16) or amp(1q21)
  - the del17 will include pts with del17 alone along with pts where the del17 is associated to t(4;14),or t(14;16) or amp(1q21)
  - the t(4;14) group will include ONLY pts with t(4;14) ALONE (no del17, t(14;16) or amp(1q21))
  - the t(14;16) group will include ONLY pts with t(14;16) ALONE (no del17, t(4;14); or amp(1q21)))
  - the amp(1q21) group will include ONLY patients with amp(1q21) ALONE (no del17, t(4;14) or t(14;16))

- Cytogenetic high-risk group defined as patients carrying any of the following cytogenetic abnormalities: del(17), translocation t(4;14), or t(14;16)
- Cytogenetic expanded high-risk group defined as: patients carrying any of the following cytogenetic abnormalities: del17, t(4;14), t(14; 16) or amp(1q21)

5.9 Pharmacokinetic and Biomarker Analysis

5.9.1 Pharmacokinetic Analyses

Plasma concentration-time data will be presented in listings. PK data will be used to perform population PK analysis using a nonlinear mixed effects modeling approach and to assess the effect of various covariates on PK after including data from other studies, if

possible. The analysis plan for the population PK analysis will be separately defined and the results of these analyses will be reported separately.

5.9.2 Biomarker Analysis

CCI

CCI

[redacted]

5.9.3 Minimal Residual Disease Analysis

The absence of minimal residual disease (MRD negativity) will be tested in all patients who achieve a CR, using bone marrow aspirates.

1) The rate of MRD negativity achieved in the ITT and CR populations will be compared between the IRd and Rd arms. PFS and OS and DOR will be assessed in MRD positive and MRD negative patients in each treatment arm.

2) The rate of MRD negativity achieved in high-risk and expanded high-risk populations will be compared between the IRd and Rd arms. PFS, OS and DOR will be assessed in MRD positive and MRD negative patients in each treatment arm.

3) PFS, OS and DOR in patients achieving MRD negativity within the first 12 months of therapy will be compared with the PFS, OS and DOR in patients achieving MRD negativity after 12 months of therapy. This will be evaluated in both arms.

5.10 Analyses of Patient-Reported Outcomes and Health Economics

5.10.1 Patient Reported Outcomes (PROs)

Patient-reported outcome (PRO) assessments using the EORTC QLQ-C30 and the MY20 will be analyzed using ITT population. The descriptive statistics of actual value and change from baseline of the subscale scores for EORTC QLQ-C30 and MY20 will be summarized by treatment group over time. Additionally, the descriptive statistics of actual values and changes from baseline of global health status/quality of life (QOL) will be summarized by treatment group over time for responders and then nonresponders. The subscales of EORTC QLQ-C30 and MY20 are defined as shown in Table 5-5 and Table 5-6.

Table 5-5 Definition of Subscale Scores of EORTC QLQ-C30

Subscale	Individual Items
Physical functioning	1-5
Role functioning	6-7
Emotional functioning	21-24
Cognitive functioning	20, 25
Social functioning	26-27
Quality of life	29-30
Fatigue	10, 12, 18
Nausea and vomiting	14-15
Pain	9, 19
Dyspnea	8
Insomnia	11
Appetite loss	13
Constipation	16
Diarrhea	17
Financial difficulties	28

Table 5-6 Definition of Subscale Scores of EORTC QLQ-MY20

Subscale	Individual Items
Future perspective	18-20
Body image	17
Disease symptoms	1-6
Side effects of treatment	7-16

Differences between treatment groups in the EORTC QLQ-C30 and MY20 subscale scores will be evaluated using published minimally important difference (MID) values. Patients with a change from baseline score ≥MID in a direction reflecting deteriorating functioning or increased symptoms at a given time point will be classified as "worsened", whereas those with a change for better of ≥MID will be classified as "improved". Those with a change from baseline score within MID will be classified as "stable". The number and percentage of patients with a change from baseline in subscale scores >=MID and <= -MID will be summarized by treatment group over time. Specific interest centers on physical functioning, global quality of life, fatigue, nausea/vomiting, pain, dyspnea, appetite loss, and constipation/diarrhea. The main endpoint for the PRO analysis will be the global health status/quality of life subscale of the EORTC QLQ-C30 and functional scales and symptom

scales of MY20. The other PRO endpoints include the remaining EORTC QLQ-C30 and MY20 subscale scores. The change from baseline in subscale scores at Cycle 18 will be presented using cumulative distribution function (CDF) figures. Additionally, CDF curves will be generated for global health status/quality of life score change from baseline at Cycle 18 by treatment group among responders and non-responders, respectively.

The change from baseline in subscale scores will be also analyzed using the repeated measures linear mixed effects models, including treatment group, baseline score, ISS stage at screening, age, sex and race as covariates. The repeated-measures analysis will use measurements collected from all available time points specified in the schedule of events in the protocol. Estimation of the variance- covariance matrix and statistics such as Akaike information criteria (AIC), Bayesian information criteria (BIC) will be included in evaluating the linear mixed-effects model. The 95% confidence intervals of the difference of the changes from baseline between the two treatments will also be provided.

Details of scoring and initial handling of missing data are included in the EORTC QLQ-C30 and MY20 scoring guidelines.

Missing data pattern will be examined. As sensitivity analyses, different imputation methods for missing data including Last Observation Carry Forward (LOCF), random slope model, and pattern mixture model may be performed if appropriate after examining missing data patterns.

5.10.2 Health Economics Analysis Using Medical Resource Utilization and Utility

CCI

[redacted]

5.10.3 Pain

CCI

[redacted]

CCI

[redacted]

Pain progression is defined as the occurrence of 1 of the following and confirmed by 2 consecutive evaluations (To qualify as progression, the patient must have a BPI-SF worst pain score ≥ 4 during pain progression):

- A ≥ 2 point and 30% increase from baseline in BPI-SF worst pain score without an decrease in analgesic use, or
- A 25% or more increase in analgesic use from baseline without a decrease in BPI-SF worst pain score from baseline

Analgesic use can be stable or increased according to the following definitions:

- Stable analgesic use is defined as less than a 25% change of the oral morphine equivalent (OME) dose from baseline
- Increased analgesic use is defined as an increase of 25% or more in OME from baseline

A sensitivity analysis will be conducted on pain progression without confirmation by 2 consecutive assessments.

In addition, the actual value and change from baseline of BPI-SF pain scores will be summarized by treatment group over time. The change from baseline in worst pain score

will be also analyzed using the repeated measures linear mixed effects models, including treatment group, baseline score, ISS stage at screening, sex, race, and age as covariates.

5.11 Safety Analyses

Safety will be evaluated by the incidence of AEs, severity and type of AEs, and by changes from baseline in the patient's vital signs, weight, and clinical laboratory results using the safety population. Exposure to the study drug regimen and reasons for discontinuation will be tabulated.

5.11.1 Adverse Events

5.11.1.1 Adverse Events

Adverse events will be coded using MedDRA. All AEs will be presented in a by-patient listing. Treatment-emergent AEs are AEs that occur after administration of the first dose of any study drug and through 30 days after the last dose of any study drug.

AEs will be tabulated according to the MedDRA by system organ class, high level terms and preferred terms and will include the following categories:

- Treatment-emergent AEs
- Drug-related treatment-emergent AEs
- Grade 3 or higher treatment-emergent AEs (also report Grade 3 and 4 separately)
- Grade 3 or higher drug-related treatment-emergent AEs (also report Grade 3 and 4 separately)
- The most commonly reported treatment-emergent AEs (ie, those events reported by ≥ 10% of patients in either treatment group)
- SAEs

Patients with the same AE more than once will have that event counted only once within each body system, once within each high level term, and once within each preferred term.

Drug-related treatment-emergent AEs will also be summarized by the National Cancer Institute Common Toxicity Criteria (NCI CTCAE) version 4.03. Patients with the same AE more than once will have the maximum intensity of that event counted within each body system, once within each high level term, and once within each preferred term.

The most commonly reported treatment-emergent AEs (ie, those events reported by ≥10% of any treatment arm) will be tabulated by preferred term. Patients with the same AE more than once will have that event counted only once within each preferred term.

An overall summary AE table will include numbers and percentages of patients who had any AE, drug-related AE, grade 3 or higher AE, grade 3 or higher drug-related AE, serious AE (SAE), drug-related SAE, AE resulting in discontinuation, and on-study deaths. On-study death is defined as the death that occurs between the first dose of any study drug and within 30 days of the last dose of any study drug.

Development of new or worsening of existing SREs (eg, new fractures, irradiation of or surgery on bone, or spinal cord compression) from baseline through the development of PD will be summarized and presented.

All concomitant medications collected from screening through the study period will be classified to preferred terms according to the World Health Organization (WHO) drug dictionary.

Two types of incidence rates will be calculated for the safety population based on the new primary malignancy assessment:

- Incidence proportions, defined as the percentage of the subjects reporting any new primary malignancy in the safety population with available information
- Incidence rates, defined by the number of the subjects reporting any new primary malignancy divided by the total duration of follow-up (patient-years = pt-yrs) in the safety population with available information up to the onset of new primary malignancies

For incidence proportions, the relative risks, defined as the ratio of incidence proportions between the 2 randomized treatment groups, were provided along with their 95% CIs. For incidence rates, the relative risks, along with their 95% CIs, will be calculated using an exponential regression model for lifetime data (assuming constant hazards).

Due to the distinct nature of hematologic and nonhematologic neoplasms, as well as the emerging signals of new primary malignancies for immunomodulating agents, analyses of new primary malignancies may be performed separately for hematologic and nonhematologic malignancies.

Additional safety analyses may be performed to most clearly enumerate rates of toxicities and to further define the safety profile of MLN9708.

Time to Resolution and Improvement of Peripheral Neuropathy Events

Peripheral neuropathy is defined as the treatment emergent adverse event in the high-level term of peripheral neuropathies NEC according to MedDRA.

A PN event is considered as resolved if its final outcome is resolved with no subsequent PN event of the same preferred term occurring on the resolution date or the day before and after. A PN event is considered as improved if the event improves from the maximum grade. That is, all the grades recorded after the maximum grade is less than the maximum grade.

Time to resolution and time to improvement are to be defined for each PN event. Time to resolution is defined as the time from the initial onset date (inclusive) to the resolution date for resolved events. Time to improvement is defined as the time from the initial onset date (inclusive) of the maximum grade to the first onset date that the toxicity grade is below the maximum grade with no higher grade thereafter, or the resolution date, whichever occurs first.

Time to improvement and time to resolution of PN events will be summarized by outcome (improvement or resolution) using the Kaplan-Meier method. The K-M survival curve and K-M medians (if estimable), along with their 2-sided 95% CIs, will be presented. This analysis is event based, thus 1 subject could contribute multiple observations if the subject has more than 1 PN event.

The analysis may be conducted for patients with any PN events or those with grade ≥ 2 PN event or those with grade ≥ 3 PN event, respectively, if data permits.

5.11.1.2 Serious Adverse Events

The number and percentage of patients experiencing at least one treatment-emergent SAE will be summarized by MedDRA primary system organ class, high level term, and preferred term. Drug-related SAE will be summarized similarly.

In addition, a by-patient listing of the SAEs will be presented (the patient listing will contain all SAEs regardless of treatment-emergent AE status).

5.11.1.3 Deaths

A by-patient listing of the deaths will be presented. All deaths occurring on-study and during follow-up will be displayed (regardless of treatment-emergent AE status).

5.11.1.4 Adverse Events Resulting in Discontinuation of Study Drug

A by-patient listing of treatment-emergent AEs resulting in discontinuation of study drug regimen will be presented.

5.11.2 Laboratory Data

For the purposes of summarization in both the tables and listings, all laboratory values will be converted to standardized units. If a lab value is reported using a non-numeric qualifier (e.g., less than (<) a certain value, or greater than (>) a certain value), the given numeric value will be used in the summary statistics, ignoring the non-numeric qualifier. However, for the bone marrow plasma cell percentage, the convention as (x-1)% (mainly for < 5% for CR) will be used.

Laboratory test results from the central laboratory will be used when they are available. Laboratory test results from local laboratory will only be used when no central laboratory test results exist at the same scheduled sample collection time point.

If a patient has repeated laboratory values for a given time point, the value from the last evaluation will be used.

Laboratory test results will be summarized according to the scheduled sample collection time point. Change from baseline will also be presented. Unscheduled laboratory test results will be listed and included in laboratory shift tables. The parameters to be analyzed are as follows:

- Hematology: hemoglobin, hematocrit, ANC, ALC, platelets, and white blood cell (WBC) count
- Serum chemistry: blood urea nitrogen, creatinine, total bilirubin, uric acid, LDH, albumin, alkaline phosphatase, AST, ALT, glucose, calcium, sodium, potassium, magnesium, phosphate, and PT.

Shift tables will be constructed for laboratory parameters to tabulate changes in NCI CTCAE for toxicity (version 4.03) from baseline to post baseline worst CTC grade. Parameters to be tabulated will include:

- Hematology: ALC, ANC, hemoglobin, platelets, WBC
- Serum chemistry: ALT, AST, alkaline phosphatase, creatinine, total bilirubin, calcium, magnesium, potassium, sodium, and phosphate.

Mean laboratory values over time through Cycle 36 for key lab parameters will be produced, including but not limited to ANC, platelets, and liver function tests (ALT/SGPT, AST/SGOT, alkaline phosphatase, and total bilirubin).

By-patient listings to be presented include hematology, serum chemistry, urinalysis, urine total protein, and urine creatinine.

5.11.3 **Electrocardiograms**

Descriptive statistics for the actual values and changes from baseline in ECGs will be tabulated by time point.

QTc interval will be calculated using Bazett's correction and Fridericia's correction, if necessary. The formulas are:

$$QTc\ (Bazett) = QT / (RR^{0.5})$$

$$QTc\ (Fridericia) = QT / (RR^{0.33})$$

where RR = 60 / heart rate (bpm)

In addition, a categorical analysis of QTc intervals will be performed for each time point. The number and percentage of patients in each QTc interval (< 450 msec, 450-480 msec, 481-500 msec, and ≥ 500msec) will be summarized at baseline and each of the subsequent time points. Categories of changes from baseline (≥ 30 msec and ≥ 60 msec) will be summarized as well.

Maximum QTc intervals and maximum changes from baseline will also be summarized similarly in a separate display.

ECG abnormalities will be presented in a data listing.

5.11.4 Vital Signs

The actual values of vital sign parameters including temperature, blood pressure, heart rate, respiratory rate, and body weight, will be summarized over time for each treatment arm. Change from baseline will also be presented.

A by-patient listing will also be presented.

5.11.5 Eastern Cooperative Oncology Group (ECOG) Performance Status

Eastern Cooperative Oncology Group performance status and change from baseline will be summarized. Shifts from baseline to the worst post-baseline score will be tabulated by treatment arm.

5.11.6 Other Safety Assessments

Pregnancy testing results will be presented in a by-patient listing.

Additional safety analyses may be performed to most clearly enumerate rates of toxicities and to further define the safety profile of MLN9708, e.g. analyses of TEAEs of clinical importance. Tables will be provided with a summary of the patient incidence of all TEAEs of clinical importance by PT, severity, and seriousness for each analysis set within each category of TEAEs of clinical importance.

6. CHANGES TO PLANNED ANALYSES FROM PROTOCOL

Reference materials for this statistical plan include Clinical Study Protocol C16014 (Protocol Amendment 3 dated 10 May 2017).

7. PROGRAMMING CONSIDERATIONS

7.1 Statistical Software

SAS version 9.1 (or higher) will be used for all analyses.

7.2 Rules and Definitions

Patient populations are defined in Section 2.

Baseline values are defined in Section 5.4.2.

8. APPENDIX

8.1 Proof of Strong Control of Type I Error Rate

Proof of strong control of Type I error rate for testing PFS and OS in ITT and PFS in subgroups:

With the proposed testing procedure for the PFS testing in ITT population and three subgroups and OS testing in ITT population, this is to prove the strong control of overall Type I error rate at one-sided 0.025 level. All alpha specified in the proof is one-sided.

We will first prove the strong control of Type I error rate under the original plan without sample size re-estimation for OS. The proof can be easily generalized to incorporate the OS sample size adaptation by switching the regular logrank test statistics at final analysis with the CHW test statistic. All the equations related to OS ITT testing still hold because the join multivariate distribution of log-rank test statistics at IA1, IA2 and FA based on planned design is the same as the log-rank test statistics at IA1, IA2, and CHW test statistic at FA.

To facilitate the probability presentation, we introduce the following notations. Let the family of null hypotheses of interest be:

- $H_0^{PFS}: S_1^{PFS}(t) = S_0^{PFS}(t)$ (no difference in PFS ITT between treatment and control arm)
- $H_0^{PFS_1}: S_{S_1,1}^{PFS}(t) = S_{S_1,0}^{PFS}(t)$ (no difference in PFS subgroup 1 between treatment and control)
- $H_0^{PFS_2}: S_{S_2,1}^{PFS}(t) = S_{S_2,0}^{PFS}(t)$ (no difference in PFS subgroup 2 between treatment and control)
- $H_0^{PFS_3}: S_{S_3,1}^{PFS}(t) = S_{S_3,0}^{PFS}(t)$ (no difference in PFS subgroup 3 between treatment and control)
- $H_0^{OS}: S_1^{OS}(t) = S_0^{OS}(t)$ (no difference in OS ITT between treatment and control arm)

Let T_1^P, T_2^P (and p_1^P, p_2^P) denote the ITT PFS logrank test statistic (and corresponding p-values) at IA1 and IA2; T_1^O, T_2^O , T_3^O (and p_1^O, p_2^O, p_3^O) denote the ITT OS logrank test statistic (and corresponding p-values) at IA1, IA2, and FA; T_{S1}, T_{S2} , T_{S3} (and p_{S1}, p_{S2}, p_{S3}) denote the PFS logrank test statistic (and corresponding p-values) at IA2 for subgroup 1, 2 and 3. Also let $p_{S(1)}$, $p_{S(2)}$, $p_{S(3)}$ denote the ordered p-values among the three subgroups; $p_{S(1)}^{\{1,2\}}$, $p_{S(2)}^{\{1,2\}}$ denote the ordered p-values among subgroup 1, and 2; $p_{S(1)}^{\{1,3\}}$, $p_{S(2)}^{\{1,3\}}$ denote the ordered p-values among subgroup 1, and 3; $p_{S(1)}^{\{2,3\}}$, $p_{S(2)}^{\{2,3\}}$ denote the ordered p-values among

subgroup 2, and 3. Let c_1, c_2 be the critical value for PFS ITT testing based on O'Brien Fleming alpha spending function where $P\{p_1^P < c_1 \text{ or } p_2^P < c_2\}$=0.02 under H_0^{PFS}; d_1, d_2, d_3 be the critical value for OS ITT testing based on O'Brien Fleming alpha spending function where $P\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\} = 0.02$ under H_0^{OS}; d_3^* be the new critical value for OS ITT testing at FA where d_3^* is calculated such that $P\{p_1^O \geq d_1, p_2^O \geq d_2, p_3^O < d_3^*\}$=0.025 - $P\{p_1^O < d_1 \text{ or } p_2^O < d_2\}$ under H_0^{OS}.

Since the key secondary endpoint - OS in ITT population is not of interest unless efficacy in the primary endpoint - PFS in ITT population is shown, there are defined paths to decision making. Liu and Hsu (2006) [6] outlined a decision path principle stating that null hypotheses should be formulated so that decision making naturally follows logical paths. We will follow this principle in formulating the null hypotheses in this proof. As a result, instead of testing all 2^5-1=31 intersection hypotheses by closed testing, we only need to test (2+1)*(2^3)-1=23 hypotheses as listed in Table 1.

Table 1: Partition hypotheses following decision paths for the proposed testing procedure

Index	Partition Hypothesis	Rejection Rule
1	$H_0^{PFS} \cap H_0^{PFS_1} \cap H_0^{PFS_2} \cap H_0^{PFS_3}$	$\{p_1^P < c_1 \text{or } p_2^P < c_2\}$ or $\{ p_{S(1)} < \frac{0.005}{3}$ or $p_{S(2)} < \frac{0.005}{2}$ or $p_{S(3)} < 0.005\}$
2	$H_0^{PFS} \cap H_0^{PFS_1} \cap H_0^{PFS_2} \cap (H_0^{PFS_3})^c$	$\{p_1^P < c_1 \text{or } p_2^P < c_2\}$ or $\{ p_{S(1)}^{\{1,2\}} < \frac{0.005}{2}$ or $p_{S(2)}^{\{1,2\}} < 0.005\}$
3	$H_0^{PFS} \cap H_0^{PFS_1} \cap (H_0^{PFS_2})^c \cap H_0^{PFS_3}$	$\{p_1^P < c_1 \text{or } p_2^P < c_2\}$ or $\{ p_{S(1)}^{\{1,3\}} < \frac{0.005}{2}$ or $p_{S(2)}^{\{1,3\}} < 0.005\}$
4	$H_0^{PFS} \cap H_0^{PFS_1} \cap (H_0^{PFS_2})^c \cap (H_0^{PFS_3})^c$	$\{p_1^P < c_1 \text{or } p_2^P < c_2\}$ or $\{ p_{S1} < 0.005\}$
5	$H_0^{PFS} \cap (H_0^{PFS_1})^c \cap H_0^{PFS_2} \cap H_0^{PFS_3}$	$\{p_1^P < c_1 \text{or } p_2^P < c_2\}$ or $\{ p_{S(1)}^{\{2,3\}} < \frac{0.005}{2}$ or $p_{S(2)}^{\{2,3\}} < 0.005\}$
6	$H_0^{PFS} \cap (H_0^{PFS_1})^c \cap H_0^{PFS_2} \cap (H_0^{PFS_3})^c$	$\{p_1^P < c_1$ or $p_2^P < c_2\}$ or $\{ p_{S2} < 0.005\}$
7	$H_0^{PFS} \cap (H_0^{PFS_1})^c \cap (H_0^{PFS_2})^c \cap H_0^{PFS_3}$	$\{p_1^P < c_1$ or $p_2^P < c_2\}$ or $\{ p_{S3} < 0.005\}$
8	$H_0^{PFS} \cap (H_0^{PFS_1})^c \cap (H_0^{PFS_2})^c \cap (H_0^{PFS_3})^c$	$\{p_1^P < c_1$ or $p_2^P < c_2\}$
9	$(H_0^{PFS})^c \cap H_0^{OS} \cap H_0^{PFS_1} \cap H_0^{PFS_2} \cap H_0^{PFS_3}$	$\{p_1^O < d_1$ or $p_2^O < d_2$ or $p_3^O < d_3\}$ or $\{ p_{S(1)} < \frac{0.005}{3}$ or $p_{S(2)} < \frac{0.005}{2}$ or $p_{S(3)} < 0.005\}$

10	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap H_0^{PFS_1} \cap H_0^{PFS_2} \cap \left(H_0^{PFS_3}\right)^c$	$\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\}$ or $\{ p_{S(1)}^{\{1,2\}} < \frac{0.005}{2} \text{ or } p_{S(2)}^{\{1,2\}} < 0.005\}$
11	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap H_0^{PFS_1} \cap \left(H_0^{PFS_2}\right)^c \cap H_0^{PFS_3}$	$\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\}$ or $\{ p_{S(1)}^{\{1,3\}} < \frac{0.005}{2} \text{ or } p_{S(2)}^{\{1,3\}} < 0.005\}$
12	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap H_0^{PFS_1} \cap \left(H_0^{PFS_2}\right)^c \cap \left(H_0^{PFS_3}\right)^c$	$\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\}$ or $\{ p_{S1} < 0.005\}$
13	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap \left(H_0^{PFS_1}\right)^c \cap H_0^{PFS_2} \cap H_0^{PFS_3}$	$\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\}$ or $\{ p_{S(1)}^{\{2,3\}} < \frac{0.005}{2} \text{ or } p_{S(2)}^{\{2,3\}} < 0.005\}$
14	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap \left(H_0^{PFS_1}\right)^c \cap H_0^{PFS_2} \cap \left(H_0^{PFS_3}\right)^c$	$\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\}$ or $\{ p_{S2} < 0.005\}$
15	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap \left(H_0^{PFS_1}\right)^c \cap \left(H_0^{PFS_2}\right)^c \cap H_0^{PFS_3}$	$\{p_1^O < d_1 \text{ or } p_2^O < d_2 \text{ or } p_3^O < d_3\}$ or $\{ p_{S3} < 0.005\}$
16	$\left(H_0^{PFS}\right)^c \cap H_0^{OS} \cap \left(H_0^{PFS_1}\right)^c \cap \left(H_0^{PFS_2}\right)^c \cap \left(H_0^{PFS_3}\right)^c$	$p_1^O < d_1$ or $p_2^O < d_2$ or $p_3^O < d_3^*$
17	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap H_0^{PFS_1} \cap H_0^{PFS_2} \cap H_0^{PFS_3}$	$p_{S(1)} < \frac{0.005}{3}$ or $p_{S(2)} < \frac{0.005}{2}$ or $p_{S(3)} < 0.005$
18	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap H_0^{PFS_1} \cap H_0^{PFS_2} \cap \left(H_0^{PFS_3}\right)^c$	$p_{S(1)}^{\{1,2\}} < \frac{0.005}{2}$ or $p_{S(2)}^{\{1,2\}} < 0.005$
19	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap H_0^{PFS_1} \cap \left(H_0^{PFS_2}\right)^c \cap H_0^{PFS_3}$	$p_{S(1)}^{\{1,3\}} < \frac{0.005}{2}$ or $p_{S(2)}^{\{1,3\}} < 0.005$
20	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap H_0^{PFS_1} \cap \left(H_0^{PFS_2}\right)^c \cap \left(H_0^{PFS_3}\right)^c$	$p_{S1} < 0.005$
21	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap \left(H_0^{PFS_1}\right)^c \cap H_0^{PFS_2} \cap H_0^{PFS_3}$	$p_{S(1)}^{\{2,3\}} < \frac{0.005}{2}$ or $p_{S(2)}^{\{2,3\}} < 0.005$
22	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap \left(H_0^{PFS_1}\right)^c \cap H_0^{PFS_2} \cap \left(H_0^{PFS_3}\right)^c$	$p_{S2} < 0.005$
23	$\left(H_0^{PFS}\right)^c \cap \left(H_0^{OS}\right)^c \cap \left(H_0^{PFS_1}\right)^c \cap \left(H_0^{PFS_2}\right)^c \cap H_0^{PFS_3}$	$p_{S3} < 0.005$

By partition principle, as long as each of the disjoint partition hypothesis is tested at level 0.025, the overall Type I error rate is also strongly controlled at the same level.

For hypotheses 17-23, Huang and Hsu (2007) [7] showed that the rejection rule in Table 1 is equivalent to the Hochberg procedure with overall 0.005 for testing the three subgroups.

For hypothesis 1, using Bonferroni inequality, the probability of false rejection is no greater than $P\{p_1^P < c_1 \text{or } p_2^P < c_2\} + P\{ p_{S(1)} < \frac{0.005}{3} \text{ or } p_{S(2)} < \frac{0.005}{2} \text{ or } p_{S(3)} <$

$0.005\}$=0.02+0.005=0.025. Similarly, for hypotheses 2-7, using Bonferroni inequality easily shows that the probability of false rejection is no greater than 0.02+0.005=0.025.

For hypothesis 8, P$\{p_1^P < c_1$ or $p_2^P < c_2\}$=0.02<0.025.

For hypothesis 9, using Bonferroni inequality, the probability of false rejection is no greater than P$\{p_1^O < d_1$ or $p_2^O < d_2$ or $p_3^O < d_3\}$ + P$\{ p_{S(1)} < \frac{0.005}{3}$ or $p_{S(2)} < \frac{0.005}{2}$ or $p_{S(3)} < 0.005\}$=0.02+0.005=0.025. Similarly, for hypotheses 10-15, using Bonferroni inequality easily shows that the probability of false rejection is no greater than 0.02+0.005=0.025. For hypothesis 16, P$\{p_1^O < d_1$ or $p_2^O < d_2$ or $p_3^O < d_3^*\}$= P$\{p_1^O < d_1$ or $p_2^O < d_2\}$ + P$\{p_1^O \geq d_1, p_2^O \geq d_2, p_3^O < d_3^*\}$+ =0.025.

Since each partition hypothesis in Table 1, is tested at 0.025 level, the overall Type I error rate is also controlled at the 0.025 level.

Next is to see that after collating results from the rejection rules in Table 1, it is equivalent to the proposed testing procedure. In order to reject H_0^{PFS}, all of partition hypotheses 1-8 have to be rejected (since they involve the null space of H_0^{PFS}) which means $\{p_1^P < c_1$ or $p_2^P < c_2\}$ which corresponds to the group sequential testing of PFS in ITT population. In order to reject $H_0^{PFS_1}$, hypotheses 1-4, 9-12, 17-20 have to be rejected which is the same as requiring hypothesis 17-20 be rejected. Similarly, hypothesis 17, 18, 21, 22 are required to be rejected for $H_0^{PFS_2}$ and hypotheses 17, 19, 21, 23 are required to be rejected for $H_0^{PFS_3}$. All the involved hypotheses are 17-23 and based on Huang and Hsu (2007), the testing procedure is exactly the Hochberg procedure with overall alpha of 0.005 level. In order to reject H_0^{OS}, all of partition hypotheses 1-16 have to be rejected. This means PFS in ITT population has to be rejected first (hypotheses 1-8). Then either $\{p_1^O < d_1$ or $p_2^O < d_2$ or $p_3^O < d_3\}$ when $p_{S(3)} \geq 0.005$; or $\{p_1^O < d_1$ or $p_2^O < d_2$ or $p_3^O < d_3^*\}$ when $p_{S(3)} < 0.005$ has to hold. This means OS in ITT population can either be rejected at first IA based on d_1 (given PFS in ITT is rejected first) or second IA based on d_2 (given PFS in ITT is rejected); if not, depending on whether all three subgroups at second IA can be rejected or not, OS in ITT population can be tested again at final analysis based on either d_3 or d_3*.

For the other two key secondary endpoints, CR rate will be tested at the same alpha level, instead of same critical value, as that for OS whenever OS reaches statistical significance. Pain response rate will be tested at the same alpha level as that for CR rate whenever CR

rate reaches statistical significance. Due to the closed sequential testing property, the family-wise error rate is strongly controlled for both the primary endpoint and three key secondary endpoints.

8.2 **Calculating the Significance Boundary for ITT PFS at IA2**

All alpha specified in this section is one-sided.

In the previous SAP (SAP version 1), the significance boundaries c_1 and c_2 for ITT PFS at IA1 and IA2 were to be calculated based on the Gamma(-1) alpha-spending approach. Specifically, c_1 was to be calculated using the observed number of PFS events at IA1 and in anticipation of 435 PFS events at IA2. c_2 was to be calculated with the purpose of exhausting the remaining available alpha while considering the correlation between c_1 and the observed number of PFS events at IA2.

This SAP (SAP version 2) modifies the target number of PFS events at IA2 from approximately 435 to approximately 370. It is also proposed after the Sponsor observed an aggregate 328 PFS events at IA1, while remaining blinded to any data by treatment arm. Therefore, to preserve the type I error rate, the significance boundary for ITT PFS at IA2 (the FA for ITT PFS) will be re-calculated while the boundary at the past IA will remain unchanged. What follows is an example calculation for the situation that exactly 370 PFS events are observed at IA2. Note that the actual value of c_2 will vary slightly depending on the eventual observed number of events.

Given 328 PFS events observed at IA1, 435 PFS events planned at IA2, and the Gamma(-1) alpha-spending function, the one-sided significance cutoffs for the log-rank test statistic T_1 and p-value p_1 at IA1 are given by u_1=2.223 and c_1=0.0131, respectively. Now suppose IA2 is performed after 370 PFS events are observed instead of the planned 435. Because the correlation between the log-rank test statistic T_2 at IA2 and T_1 changes as a result, the information fraction I at IA1 should be adjusted to exhaust the remaining alpha while preserving the type I error rate.

Set I=328/370 and keep u_1=2.223. That is, update the information rate at IA1, but fix the alpha already spent at IA1. Under the null hypothesis of no treatment benefit, the vector (T_1, T_2) follows the bivariate normal distribution

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{I} \\ \sqrt{I} & 1 \end{pmatrix} \right)$$

By performing a grid search, it can be found that setting u_2=2.131 preserves type I error rate at alpha-level 0.02:

$$\Pr(T_1 > u_1) + \Pr(T_1 < u_1 \text{ and } T_2 > u_2) = 0.02$$

The p-value cutoff corresponding to u_2 is c_2=0.0165.

9. REFERENCES

1) Durie BG, Harousseau JL, Miguel JS, Blade J, Barlogie B, Anderson K, et al. International uniform response criteria for multiple myeloma. Leukemia 2006; 20(9): 1467-73.

2) Lan K, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-63.

3) Robins JM, Hernan M, Brumback B. Marginal strucural models and causal inference in epidemiology. Epidemiology 2000; 11(5): 550-560.

4) Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000;56(3):779-88.

5) D'Agostino R.B., Lee M.L., Belanger A.J., Cupples L.A., Anderson K., and Kannel W.B. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. Statistics in Medicine, 9(12): 1501–1515, 1990.

6) Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; 55:853–857.

7) Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group sequential trials. *Stat Med* 2010;29(2):219-28.

8) Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84

9) Liu Y, Hsu J (2009). Testing for efficacy in primary and secondary endpoints by partitioning decision paths. *Journal of the American Statistical Association*, 104: 1661-1670.

10) Huang Y, Hsu J; Hochberg's Step-Up Method: Cutting Corners Off Holm's Step-Down Method. *Biometrika* 2007; 94 (4): 965-975. doi: 10.1093/biomet/asm067

ELECTRONIC SIGNATURES

Signed by	Meaning of Signature	Server Date (dd-MMM-yyyy HH:mm 'UTC')
PPD	Pharmacovigilance Approval	14-Jan-2020 19:52 UTC
	Clinical Science Approval	14-Jan-2020 19:56 UTC
	Clinical Approval	14-Jan-2020 19:56 UTC
	Biostatistics Approval	15-Jan-2020 14:39 UTC