Title: A Randomized Phase III Study Comparing Conventional Dose Treatment Using a Combination of Lenalidomide, Bortezomib and Dexamethasone (RVD) to High-Dose Treatment with Peripheral Stem Cell Transplant in the Initial Management of Myeloma in Patients up to 65 Years of Age.

NCT Number: NCT01208662

IRB Approval Date: 07/12/2023

DFCI 10106 Statistical Analysis Plan

Protocol Title: (DFCI: 10106)	A Randomized Phase III Study Comparing Conventional Dose Treatment Using a Combination of Lenalidomide, Bortezomib and Dexamethasone (RVD) to High-Dose Treatment with Peripheral Stem Cell Transplant in the Initial Management of Myeloma in Patients up to 65 Years of Age
Protocol Date:	November 30, 2016
SAP Authors:	[redacted]
SAP Version:	9
SAP Date:	November 2021
SAP Changes	November 2021: Modified Table 1 as follows based on the approach used for DMC analyses- Removed the withdrawal rows due to duplication with other categories; Added footnote to clarify that if death occurs beyond 1 year of last disease evaluation then censored at date of last disease evaluation (based on study calendar where follow-up after 3 years from last patient entered, additional follow-ups are according to institutional practice); Deaths without PD date are counted as events for PFS, EFS and censored for TTP at date of death; Added footnote to clarify that subjects who start non-protocol therapy prior to the last disease evaluation fall under the initiation of new treatment category and are censored at the date of the initiation of new treatment or date treatment ended if non-protocol therapy date is unknown; Added footnote to clarify that if new treatment within 1 month prior to progression, patients were coded as having an event at time of progression; For non-protocol therapy, the date used in the analysis was modified from date treatment ended to date of initiation of new treatment; and Added rules for event-free survival; August 1, 2019: Susanna Jacobus to replace Bhup Rawal as MS statistician. March 19, 2019: Added 3rd paragraph to section 5.1 and 2nd paragraph to section 12.1; Modified Table 1 to clarify coding of

progressions; April 10, 2018: Updated elements of tables to be consistent with the study; October 2017: Editorial changes;October 2016: Increased accrual to 720 patients; September 2015: Changed MS statistician; July 2013: Major changes required after the separation of the IFM and DFCI Studies. The original SAP date with combined analysis of the two studies is November 2011.

TABLE OF CONTENTS

1. Preface

This statistical analysis plan (SAP) describes the planned analysis and reporting for DFCI 10106 protocol A Randomized Phase III Study Comparing Conventional Dose Treatment Using a Combination of Lenalidomide, Bortezomib and Dexamethasone (RVD) to High-Dose Treatment with Peripheral Stem Cell Transplant in the Initial Management of Myeloma in Patients up to 65 Years of Age (IFM/DFCI 2009).

For fifteen years, high-dose therapy (HDT) has been the standard treatment for multiple myeloma (MM) in younger patients. In the 1990s, several randomized studies demonstrated the superiority of high-dose treatments versus conventional chemotherapies in terms of response, event-free survival and overall survival (OS). The superiority of HDT over conventional-dose therapy is related to obtaining a higher rate of very good partial response (VGPR) or better, which in turn is correlated with longer PFS, but only in some studies with OS. Indeed, a recent meta-analysis by Koreth and colleagues demonstrated PFS and no OS advantage.

For the last 4-5 years, the arrival of novel therapies (thalidomide, bortezomib and lenalidomide) has revolutionized conventional therapeutic regimens. The use of these new therapies has improved complete response (CR) and VGPR rates of HDT as well as those of conventional-dose therapy, to such a point that these rates have now become similar in both groups of treatment. Thus, the arrival of novel therapies has brought into question the necessity of HDT as first-line therapy in young patients.

This phase III study is being completed to compare the efficacy, quality of life and cost of high dose therapy to those of conventional-dose treatment, with both treatment arms receiving novel drugs as part of induction, consolidation and maintenance in myeloma patients up to 65 years of age.

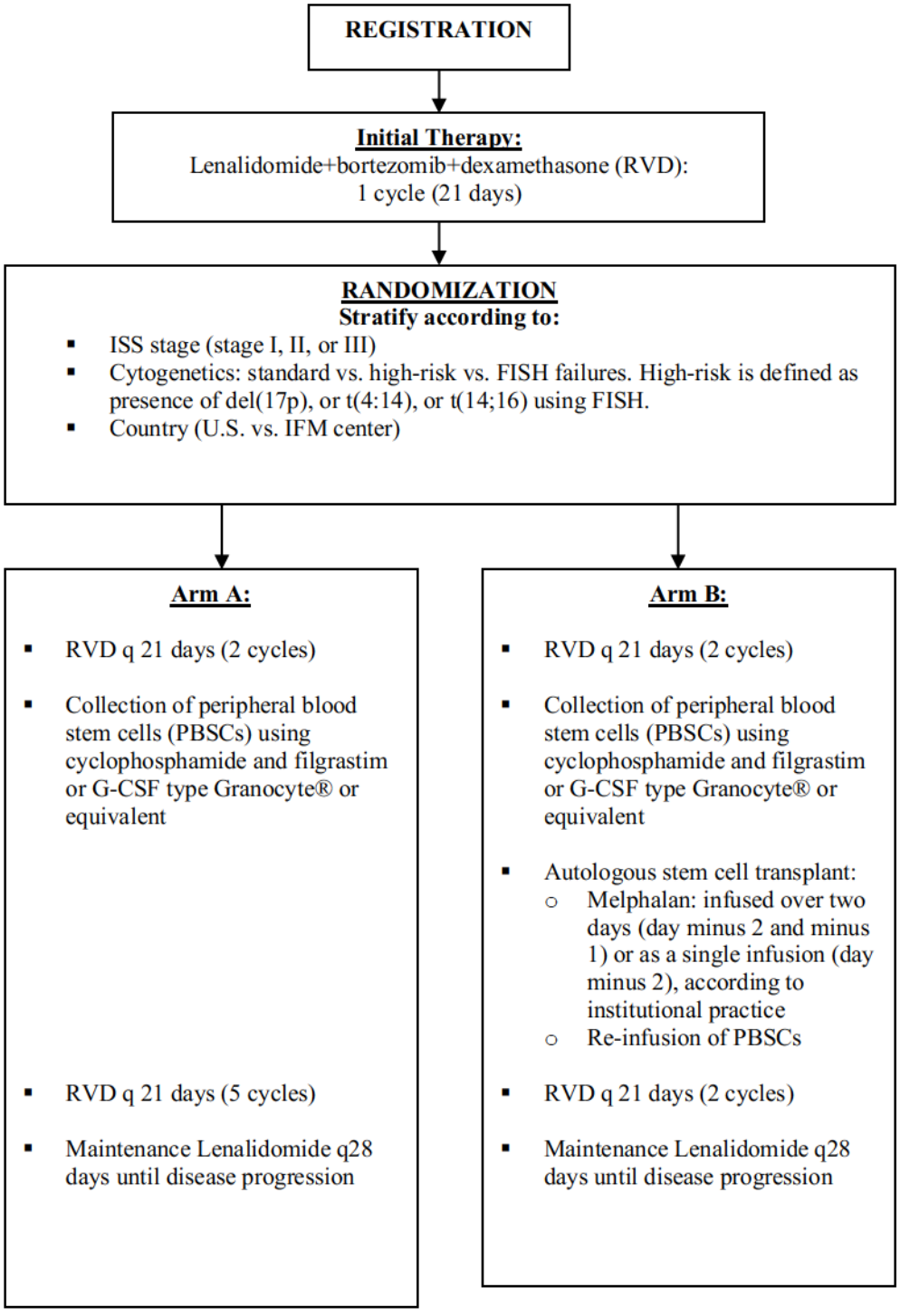
The structure and content of this SAP was designed to follow the guidelines in the International Conference on Harmonisation of Technical Requirements for the Registration of Pharamceuticals for Human Use (ICH); Guidance on Statistical Principles in Clinical Trials [1] . All work planned and reported for this SAP will follow internationally accepted guidelines, published by the American Statistical Association[2] and the Royal Statistical Society[3] , for statistical practice.

The following documents were reviewed in preparation for this SAP:

- Clinical Research Protocol (DFCI 10106) and amendments
- Electronic case report forms
- ICH Guidelines on Statistical Principles for Clinical Trials

The reader of this SAP is encouraged to also read the clinical research protocol for details on the conduct of this study, and the operational aspects of clinical assessments and timing for completing a patient in this study.

1.1 Study Schema



2. Purpose of SAP

The purpose of this SAP is to outline the planned analysis to be completed for the Clinical Study Report (CSR) for protocol DFCI 10106. The planned analyses indentified in this SAP will be included in regulatory submissions if appropriate and future manuscripts. Also, exploratory analyses not necessarily identified in this SAP may be performed to support the clinical development program (such as research in a program project). Any post-hoc, or unplanned, analyses not identified in this SAP performed will be clearly identified as such in the Clinical Study Report.

3. Study Objectives and Endpoints

3.1 Primary Objectives

To compare progression-free survival (PFS) between Arm A and Arm B

3.2 Secondary Objectives

- To compare the response rates (RR) between the two arms
- To compare time to progression (TTP) between the two arms
- To compare the overall survival (OS) between the two arms
- To compare the toxicity between the two arms
- To define genetic prognostic groups evaluated by gene expression profiling (GEP)
- To examine the best treatment in each Gene Expression Profile-defined prognostic group
- To compare quality of life (QOL) between the two arms
- To collect medical resource utilization (MRU) information which may be used in economic evaluation models. This objective will be addressed in the US sites only.

3.3 Primary Endpoint Definitions

Progression-Free Survival (PFS): the primary endpoint in this study. PFS is defined as the time from randomization to the disease progression or death from any cause. Patients who have not progressed or died are censored at the date last known progression-free.

3.4 Secondary Endpoint Definitions

Response Rates and Duration of Response : The disease response will be assessed using criteria based on the International Working Group Uniform Response Criteria (IMWG Appendix 2). If the only measurable parameter is serum immunoglobulins free light chain (FLC), the participant will be

followed by FreeLite™ Disease Response Criteria provided in Appendix 2. Disease response by the Modified EBMT Response Criteria will also be collected on participants as a secondary measure. The same method of assessment and technique will be used for disease measurement at baseline and during follow-up. Disease response should be confirmed by two consecutive assessments made at anytime before the initiation of any new therapy by the IMWG criteria and at a minimum of 6 weeks apart by the EBMT criteria.

Central review of disease response assessment is planned for this trial. Central review will be performed on the following disease response measures: M-protein quantification and immunofixation from serum and 24-hour urine collection and serum freelite testing. Results from the central review of response will be recorded in a separate database and will be compared with the response data in the SAS dataset in the final analysis.

The duration of overall response is measured as the time from initiation of first response to first documentation of disease progression or death. Patients who have not progressed or died are censored at the date last known progression-free. The duration of overall CR is measured as the time from initiation of CR to first documentation of disease progression or death. Patients who have not progressed or died are censored at the date last known progression-free.

Time to progression: Time to progression is defined as the time of randomization until progression. Patients who have died without evidence of progression are censored in the TTP analysis at the time of death and patients who are alive without progression are censored at the last disease assessment.

Overall survival (OS): OS is defined as the time from randomization to death. Alive patients are censored at the date last known alive.

Toxicity: Descriptions and grading scales found in the CTCAE version 4.0 NCI Common Terminology Criteria for Adverse Events are used. Abnormal laboratory values or diagnostic test results constitute adverse events only if they induce clinical signs or symptoms or require treatment or further diagnostic tests. Safety assessments will be collected during the treatment emergent period, defined as the time from initiation of study treatment up to 30 days last of dose of study drug or the date of start of investigational agent. Safety endpoints include any adverse event (expected and unexpected), serious adverse event (SAE), lethal toxicities, secondary malignancy, laboratory safety assessment, ECOG, ECGs and vital signs. Duration, intensity, attribution and time to onset of toxicities will be collected.

QOL Endpoints : Three QOL instruments will be evaluated in this study: The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire - Core (QLQ-C30), the EORTC QLQ-MY20 Multiple Myeloma module, and the Functional Assessment of Cancer Therapy/Gynecologic Oncology Group-Neurotoxicity (FACT-NTX) side-effects questionnaires. QOL domains will be compared between Arm A and Arm B, and include health related quality of life, distress, psychological functioning, physical well-being and functional well-being.

Medical Resource Utilization (Pharmacoeconomic) Endpoints (US Sites only):
Pharmacoeconomics assessment will measure the costs of the two treatment arms by comparing

markers of resource utilization between Arm A and Arm B. Medical resource utilization (MRU) data associated with medical encounters related to disease or myeloma therapies will be collected. Specifically, MRU is evaluated based on the number of medical care encounters such as hospital admissions and their duration, outpatient visits, and diagnostic tests and procedures. The MRU data will be used to conduct economic analyses.

The EQ-5D will be used to capture participant-reported utilities for use in cost-utility analyses in this trial. The EQ-5D contains a five item survey with three response levels per item measuring mobility, self-care, usual activities, pain/discomfort and anxiety/depression. It also contains a 0-100 visual analogue scale to measure current overall health and 9 optional socio-demographic questions.

4. Study Methods

4.1 Overall Study Design and Plan

The primary objective of this protocol is to compare PFS of Arms A (conventional dose arm) and B (high dose arm). The primary analysis of PFS will be performed using a stratified two-sided log-rank test with an overall type I error rate of 5% on the intent-to-treat population consisting of all randomized patients. Cases determined to be ineligible after randomization will be included in the analysis. PFS in each of the arms will be estimated using the method of Kaplan and Meier .

4.1.1 Summary of Design Changes

In the original design, data from this trial and a parallel trial in France were to be combined together for analysis to achieve the objectives of this study. The original protocols for both studies included maintenance lenalidomide for one year. Based on the NEJM maintenance paper (CALGB NEJM 2012) reporting the benefit of maintenance therapy the US protocol was revised October 16, 2012 to extend maintenance until progression. The accrual for the US was expanded to include 660 patients to address the question of conventional dose followed by maintenance until PD versus high dose therapy arm followed by maintenance until PD. The accrual for the IFM protocol was 700 patients to address the question of conventional dose followed by 1-year of maintenance vs. high dose therapy arm followed by 1-year of maintenance.

In February 2016, McCarthy et al presented results from a meta-analysis showing a 50% reduction in the hazards for continuous maintenance therapy (EHA February 2016). With a reduction in the failure rate, the time to the full information could be longer than expected in this study. Therefore, to account for this potential reduction in the hazard rates in both arms (the hazard ratio remains at 1.43), the sample size is increased to 720 randomized patients. With the increase from 660 to 720 randomized patients, the time to reach full information is 5 months earlier.

Details on the design changes are summarized below. The sections below include the original design (4.1.2), the October 16, 2012 design at the time the US maintenance treatment duration was modified (4.1.3), and the October 2016 change to increase the sample size to account for potential reduction in the hazard rates per arm (4.1.4).

4.1.2 Original Design

The IFM and DFCI studies combined had 92% power to detect a 23% reduction in the PFS hazard from 0.0231/month on conventional dose arm (Arm A) to 0.0177/month on high dose therapy arm

(Arm B) using a stratified two-sided log-rank test with an overall type I error rate of 0.05. This corresponds to a hazard ratio (hazard of conventional dose Arm A/hazard of high dose Arm B) of 1.3. Full information under the alternative hypothesis is 658 failures. Assuming median PFS of 30 months on the RVD alone arm (Arm A) and the PFS time follows an exponential distribution, this difference corresponds to a 30% increase in median survival to 39 months for Arm B. Based on these medians and corresponding failure rates, the required number of failures will be observed with 1000 patients enrolled over 36 months with 36 months of follow-up for total study time of 72 months. Two interim analyses were planned at approximately 33% and 69% information. In calendar times these are anticipated to be at approximately 30 months (prior to the end of enrollment) and 48 months with the final analyses planned at approximately 72 months. To preserve the overall type I error rate, critical values at the interim analysis will be determined using the Lan-DeMets error spending rate function corresponding to the O'Brien Fleming boundary. The study will also be monitored for early stopping in favor of the null hypothesis using Jennison-Turnbull repeated confidence interval (RCI) methodology. At each interim analysis, the one-sided 97.5% repeated confidence upper limit on the hazard ratio will be computed using the critical value from the error spending function. If the upper limit lies below the target alternative hazard ratio (hazard of the conventional dose /hazard of the high dose) then the DMC may consider stopping the trial early in favor of the conventional dose arm.

4.1.3 Design modification for US study change to continuous maintenance therapy (October 16, 2012)

With the separation of the IFM/DFCI study into two individual studies, the US study is designed with 90% power to detect a 30% reduction in the PFS hazard from 0.0231/month on conventional dose arm (Arm A) with maintenance until PD to 0.0162/month on high dose therapy arm (Arm B) with maintenance until PD. A stratified two-sided log-rank test will be used with an overall type I error rate of 0.05. This corresponds to a hazard ratio (hazard of conventional dose Arm A/hazard of high dose Arm B) of 1.43. Full information under the alternative hypothesis is 329 failures. Assuming median PFS of 30 months on the RVD alone arm (Arm A) and the PFS time follow an exponential distribution, this difference corresponds to a 43% increase in median survival to 43 months for Arm B. A power of 90% was used in this study to adjust for the potential for cross over from the conventional dose arm to the high dose therapy arm prior to progression. Based on simulations allowing for constant cross over as well as varying patterns of cross over the power is reduced approximately 7-10% with up to 15% cross over at 3 years.

It is anticipated that the US accrual will reach 100 randomized patients by 24 months. Based on the medians and the corresponding failures rates as well as the extended follow-up among the initial 100 patients, the required number of failures will be observed with an additional 560 patients entered over 30 months with 18 months of follow-up for a total study time of 72 months (24 months accrual time for the initial 100 patients, 30 months accrual time for the additional 560 patients and 18 months of follow-up). Specifically, with 560 patients entered over 30 months (monthly accrual rate of 18-19 patients/month) with 18 months of follow-up approximately 261 events will be observed assuming an exponential distribution with the above specified medians and failure rates. Among the initial 100 patients, the minimum follow-up is extended from 36 months to 48 months and therefore, based on the exponential distribution approximately 68 failures are expected. Two interim analyses will occur at 33% and 69% information and the final analysis at full information.

These results will be presented to the data-monitoring committee (DMC). At each DMC meeting toxicity results will be presented and interim analyses if at designated information times. To preserve the overall type I error rate, critical values at the interim analysis will be determined using the Lan-DeMets error spending rate function corresponding to the O'Brien Fleming boundary. The study will also be monitored for early stopping in favor of the null hypothesis using Jennison-Turnbull repeated confidence interval (RCI) methodology. At each interim analysis, the one-sided 97.5% repeated confidence upper limit on the hazard ratio will be computed using the critical value from the error spending function.

4.1.3.1 Impact of the US study design changes on the IFM study design

In the original design, the IFM and DFCI studies combined have 92% power to detect a 23% reduction in the PFS hazard from 0.0231/month on conventional dose arm (Arm A) to 0.0177/month on high dose therapy arm (Arm B) using a stratified two-sided log-rank test with an overall type I error rate of 0.05. Full information is 658 failures. Current accrual for the IFM protocol has been faster than expected with 700 randomized patients entered within 24 months. To maintain at least 80% power, patients will continue to be followed until 72 months after the first randomization with a minimum follow-up of 48 months. This would result in 489 failures under the alternative (81% power). Two interim analyses will occur at 33% and 69% information and the final analysis at full information. To preserve the overall type I error rate, critical values at the interim analysis will be determined using the Lan-DeMets error spending rate function corresponding to the O'Brien Fleming boundary. The study will also be monitored for early stopping using Jennison-Turnbull repeated confidence interval (RCI) methodology. At each interim analysis, the two-sided repeated confidence limits on the hazard ratio will be computed using the critical value from the error spending function.

4.1.4 Design modification for US study change to increase sample size to 720 patients (October 12, 2016 date)

Reason for design modifications: In February 2016, McCarthy et al presented results from a meta-analysis showing a 50% reduction in the hazards for continuous maintenance therapy (EHA February 2016) . With a reduction in the failure rate, the time to the full information could be longer than expected. Therefore, to account for this potential reduction in the hazard rate in both arms (the hazard ratio remains at 1.43) and reduce the time to full information by 5 months, the sample size is increased to 720 randomized patients. With 660 and 720 patients the total study time with the reduction in hazards is 113 and 118 months, respectively. To derive the number of months required to reach full information, the actual accrual patterns through August 2016 and projections for the remaining number of patients to reach full accrual are used in the calculations.

Study Design: This study is designed with 90% power to detect a 30% reduction in the PFS hazard of conventional dose arm (Arm A) with maintenance until PD to high dose therapy arm (Arm B) with maintenance until PD. A stratified two-sided log-rank test will be used with an overall type I error rate of 0.05. This corresponds to a hazard ratio (hazard of conventional dose Arm A/hazard of high dose Arm B) of 1.43. Full information under the alternative hypothesis is 329 failures. A power of 90% was used in this study to adjust for the potential for cross over from the conventional dose arm to the high dose therapy arm prior to progression. Based on simulations allowing for

constant cross over as well as varying patterns of cross over the power is reduced approximately 7-10% with up to 15% cross over at 3 years.

The current accrual patterns are used to determine the time required to reach full information with 720 patients randomized. Table 4.1 lists the actual accrual through August 2016 and the projected number of patients in the remaining months to reach 720 patients. Based on the failures rates (λ_A =0.0231/2= 0.01155, λ_B =0.0162/2= 0.0081), exponential distribution, and the accrual pattern in Table 14.1, the required number of failures (329) will be observed at 33 months after the end of accrual (80 months) for a total study time of 113 months. This is 5 months earlier in time than when full information would be observed with 660 patients.

Two interim analyses will occur at 33% and 69% information and the final analysis at full information. These results will be presented to the data- monitoring committee (DMC). At each DMC meeting toxicity results will be presented and interim analyses if at designated information times. To preserve the overall type I error rate, critical values at the interim analysis will be determined using the Lan-DeMets error spending rate function corresponding to the O'Brien Fleming boundary. The study will also be monitored for early stopping in favor of the null hypothesis using Jennison-Turnbull repeated confidence interval (RCI) methodology. At each interim analysis, the one-sided 97.5% repeated confidence upper limit on the hazard ratio will be computed using the critical value from the error spending function.

Proposed analysis for both the IFM and US Studies: A descriptive analysis will be performed to attempt to address a question comparing PFS and OS for maintenance lenalidomide for 1 year (700 patients) vs. maintenance lenalidomide until PD (720 patients) overall and by conventional dose (350 vs. 360 patients) /high dose therapy arm (350 vs. 360 patients).

Table 4.1. Actual Accrual Per Month from October 2010 through August 2016. Numbers for September 2016 to the June 2017 are projected numbers.

Month	Accrual	Month	Accrual	Month	Accrual
2010-10	1	2013-01	10	2015-04	9
2010-11	1	2013-02	13	2015-05	13
2010-12	3	2013-03	7	2015-06	15
2011-01	0	2013-04	11	2015-07	16
2011-02	3	2013-05	5	2015-08	14
2011-03	4	2013-06	4	2015-09	14
2011-04	1	2013-07	6	2015-10	15
2011-05	2	2013-08	11	2015-11	18
2011-06	2	2013-09	15	2015-12	10
2011-07	3	2013-10	10	2016-01	8
2011-08	7	2013-11	5	2016-02	16
2011-09	4	2013-12	11	2016-03	20
2011-10	6	2014-01	9	2016-04	10
2011-11	4	2014-02	14	2016-05	17
2011-12	2	2014-03	9	2016-06	11
2012-01	3	2014-04	7	2016-07	8
2012-02	6	2014-05	10	2016-08	7
2012-03	4	2014-06	13	2016-09*	10
2012-04	5	2014-07	18	2016-10*	10

2012-05	10	2014-08	14	2016-11*	10
2012-06	4	2014-09	11	2016-12*	10
2012-07	4	2014-10	11	2017-1*	10
2012-08	7	2014-11	14	2017-2*	10
2012-09	8	2014-12	16	2017-3*	10
2012-10	10	2015-01	13	2017-4*	10
2012-11	6	2015-02	13	2017-5*	8
2012-12	4	2015-03	17		

*Projected accrual

4.2 Data Collection and Database

There is one eCRF system used for the study with a different set of screens for the US and IFM sites due to language differences and pre-specified different data items (for example, the pharmacoeconomic secondary objective is for the US sites only and therefore, this information is collected only for the US sites). The eCRF system is the one used by the IFM and managed by a company called Statitec. The eCRFs data items and the data check specifications for this study were jointly developed with the IFM and the DFCI study team members. Changes to the data fields require the approval of IFM and DFCI study team members. SAS data set specifications have been defined for data transfer to the study statisticians. There is no change on the the eCRF system following the separation of the two studies.

4.3 Selection of Study Population

Key inclusion and exclusion criteria are identified from the Participant Selection Section of the protocol. All laboratory assessments must be within 21 days of initiation of protocol therapy.

Inclusion Criteria for Registration include:

- Participants must have a diagnosis of MM, according to International Myeloma Foundation 2003 Diagnostic Criteria.
- Participants must have documented symptomatic myeloma, with organ damage related to myeloma
- Participants must have myeloma that is measurable by either serum or urine evaluation of the monoclonal component or by assay of serum free light chains.
- Age between 18 and 65 years at the time of signing the informed consent form.
- ECOG performance status $\leq$2 (Karnofsky $\geq$60%).
- Negative HIV blood test within 21 days of study entry.
- Females of childbearing potential must have a negative serum or urine pregnancy test with a sensitivity of at least 25 mIU/mL 10 to 14 days prior to therapy and repeated again within 24 hours of starting lenalidomide and must either commit to complete abstinence from heterosexual contact or begin two acceptable methods of birth control, one highly effective method and one additional effective (barrier) method, at the same time, at least 28 days before she starts taking lenalidomide.

Exclusion Criteria for Registration Include:

- Participant treated with any prior systemic therapy with exception of localized radiotherapy and corticosteroids as specified in protocol.
- Primary amyloidosis (AL) or myeloma complicated by amylosis
- Participants receiving any other investigational agents.
- Participants with known brain metastases.
- Poor tolerability or known allergy to any of the study drugs or compounds of similar chemical or biologic composition.
- Participants with inadequate platelet level, absolute neutrophil count, hemoglobin level, hepatic function, renal function or respiratory function as defined per protocol.
- Participant with clinical signs of heart or coronary failure, or evidence of left ventricular ejection fraction (LVEF) < 40%. Participant with myocardial infarction within 6 months prior to enrollment or have New York Heart Association (NYHA) Class III or IV heart failure (see Appendix VII), uncontrolled angina, severe uncontrolled ventricular arrhythmias, or electrocardiographic evidence of acute ischemia or active conductive system abnormalities. Prior to study entry, any ECG abnormality at screening has to be documented by the investigator as not medically relevant
- Intercurrent illness including, but not limited to ongoing or active severe infection, known (active or not) infection with hepatitis B or C virus, poorly controlled diabetes, severe uncontrolled psychiatric disorder or psychiatric illness/social situations that would limit compliance with study requirements.
- Female participants pregnant or breast-feeding.
- Inability to comply with an anti-throembotic treatment regimen (e.g., administration of aspirin, enoxaparin, or low molecular weight heparin administration (type Innohep® or equivalent)
- Peripheral neuropathy $\geq$ Grade 2 on clinical examination, within 21 days of initiation of protocol therapy.

Eligibility and Exclusion Criteria for Randomization:

After registration and prior to randomization, participants will receive 1 cycle of RVD. Participants are not required to meet additional eligibility or exclusion criteria prior to randomization procedures.

There are no additional screening test requirements for randomization. However, results of cytogenetics by FISH, and beta-2 microglobulin and albumin from registration screening tests are required in order to proceed with randomization because these laboratory results are stratification factors in the randomization. The Investigator is responsible for keeping a record of the reason(s) that participants do not proceed to randomization. This information will be collected in the database.

4.4 Method of Treatment Assignment and Randomization

Randomization must occur 2-3 weeks after the initiation of cycle 1 of RVD and prior to cycle 2 of RVD. Stratified permuted blocks will be used in the randomization using the following stratification factors.

1. ISS Stage I vs. II vs. III. Institutions will provide Beta2- microglobulin level (mg/L) and serum albumin level (g/dL) entered from screening visit, and the randomization system will compute the ISS stage.
2. Standard vs. high risk vs. FISH failures. High-risk is defined as the presence of del(17p), or t(4:14), or t(14;16) using FISH
3. Country (US vs. IFM)

The reason why patients are not randomized will be documented and collected on the database. If a patient is randomized and does not receive randomized treatment, baseline and follow-up data will still be collected.

4.5 Treatment Blinding

Treatment could not be blinded in this study in which one arm contains high dose therapy.

4.6 Data Monitoring Committee and Steering Committee

A data safety and monitoring committee (DMC) and Steering Committee have been set up for this study. The DMC will meet two times per year. The DMC committee consists of 3 medical oncologists (1 from US, 2 from Europe), 2 statisticians (2 from US) and 1 lay member. The Steering Committee includes study co-chairs, two senior investigators from DFCI, two senior investigators from the IFM, one independent investigator from the US, one independent investigator from the Europe, the US study statistician, and representatives from the companies. Per the DMC Charter, the DMC committee is an advisory committee to the IFM/DFCI Steering Committee. Responsibilities of the DMC include reviewing interim toxicity data and proposing corrective actions as deemed necessary, reviewing interim analysis outcome data, reviewing proposed major changes to protocols, evaluating impact of independent scientific investigations, and reviewing requests for release or use of study outcome data. The primary responsibilities of the Steering Committee are twofold. First, the Steering Committee is responsible for maintaining the scientific integrity of the trial, for example, by recommending changes to the protocol in light of emerging clinical or scientific data from other trials. Second, the Steering Committee is responsible for translation of recommendations of the Data and Safety Monitoring Committee into decisions

5. Sequence of Planned Analyses

5.1 Interim Analysis

Interim analyses will be performed at 33% information (approximately 108 failures) and 69% information (approximately 227 failures) and the full analysis will be performed at 100% information (329 failures). To preserve the overall type I error rate, critical values at the interim analysis will be determined using the Lan-DeMets error spending rate function corresponding to the O'Brien Fleming boundary. The O-Brien Fleming upper boundary at 33%, 69%, and 100% information are 3.7330, 2.4651, and 1.9998 with corresponding normal significance levels of 0.0000946, 0.0068493, and 0.0227634 respectively. These results will be presented to the data-monitoring committee (DMC). At each DMC meeting accrual and safety information will be presented and interim analyses if at designated information times. In addition, the median PFS will be reported for the conventional dose Arm A to determine if the failure rate assumed for this group is appropriate.

The study will also be monitored for early stopping in favor of the null hypothesis using Jennison-Turnbull repeated confidence interval (RCI) methodology. At each interim analysis, the two-sided repeated confidence limits on the hazard ratio will be computed using the critical value from the error spending function.

At the time of the interim analysis, if a case does not have final status from the response review committee the institutions assessment of response per the IMWG will be used in the analysis and will follow Table 1 for coding based on the institutions assessment.

5.2 Final Analysis and Reporting

All final, planned, analyses identified in the protocol and in this SAP will be performed only after the full information is achieved. Any post-hoc, exploratory analyses completed to support planned study analyses, which were not identified in this SAP, will be documented and reported in the clinical study report as such. Any results from these unplanned analyses will also be clearly identified in the text of the clinical study report.

6. Sample Size Determination

This study is designed to have adequate power to detect a 30% reduction in the PFS hazard rate and a hazard ratio (hazard of conventional dose Arm A/hazard of high dose Arm B) of 1.43. The primary analysis of PFS will be performed using a stratified two-sided log-rank test with an overall type I error rate of 5%. A power of 90% was used in this study to adjust for the potential for cross over from the conventional dose arm to the high dose therapy arm prior to progression. Based on simulations allowing for constant cross over as well as varying patterns of cross over the power is reduced approximately 7-10% with up to 15% cross over at 3 years.

7. Analysis Populations

The following analysis populations are planned for the studies:

Intent-to-treat (ITT) population: The ITT population is defined as all patients who are randomized to the study drug in this study. Patients in this population will be analyzed according to the treatment arm they are assigned by randomization, regardless of treatment actually received or any dosing error. Patients who complete 1 cycle of therapy but do not proceed to randomization will not be included in this population. The ITT population will be used for all efficacy analysis.

Safety population: The safety population is defined as all patients who are randomized. Patients who complete 1 cycle of initial therapy but do not proceed to randomization will be included in this population, but will be identified separately as they are not randomized. The safety population will be used to conduct safety analysis.

As-treated population (ATP): The as-treated population is classified by the actual treatment they receive after randomization and is based on all patients who receive at least one dose of study drug. Patients who complete 1 cycle of initial therapy but do not proceed to randomization will not be included in the as-treatment analysis. The ATP population will be used as a sensitivity analysis to check the robustness of the ITT results.

8. General Issues for Statistical Analysis

8.1 Overview

Patient characteristics will be summarized using proportions for discrete data and median for continuous variables with comparisons between arms performed using Fisher's exact test for discrete data and Mann-Whitney rank sum for continuous data. Time to event outcomes will be estimated using Kaplan-Meier methods and compared between groups using stratified log rank tests. Logistic regression models and Cox proportional hazard regression models will be implemented to evaluate the impact of baseline information on response and time to event outcomes. Evaluation of factors associated with better outcome will include characteristics of the patient and of the myeloma, biological and genetic markers. It will be performed using the same statistical tests (log rank and Cox model).

The final efficacy and safety analysis will take place when approximately 329 PFS events are reached. For the primary PFS analysis, a stratified two-sided log-rank test with a cumulative type I error rate of 0.05 will be used. No multiplicity will be considered in the following secondary endpoints (response rate, duration of response, time to progression, overall survival) and any p-value that is ≤ 0.05 will be identified as nominally significant. Longitudinal methods will be used to evaluate the quality of life endpoints. Missing data is anticipated for quality of life endpoints, and therefore, multiple imputation methods will be applied for the primary analysis of quality of life endpoints and the complete case analysis will be considered as a sensitivity analysis.

8.2 Analysis Software

All analysis will be performed using SAS Software version 9.1 or later (SAS Institute INC, Carey, NJ) and the R package.

8.3 Methods for Withdrawals and Missing Data

Primary efficacy analysis of PFS is performed on the intent to treat population. All patients will be followed until progression, per protocol, regardless of reason for study withdrawal, unless the patients withdraw consent and ask not to be followed for progression.

Missing data is anticipated for the quality of life endpoints. To account for this, multiple imputation methods will be applied for the primary analysis of quality of life endpoints and the complete case analysis will be considered as a sensitivity analysis.

8.4 Multicenter Studies

Multicenter issues related to study management and data collection and database are provided in sections 4.1 and 4.2.

8.5 Multiple Comparisons and Multiplicity

Adjustments for interim analysis have been incorporated into the sequential monitoring plan.

No multiplicity will be considered in the following secondary endpoints (response rate, duration of response, time to progression, overall survival) and any p-value that is ≤ 0.05 will be identified as nominally significant. Longitudinal methods will be used to evaluate the quality of life endpoints,

however, Bonferroni procedure will be used to account for multiple comparisons if results for each assessment time are reported individually.

8.6 Planned Subgroups, Interactions and Covariates

Subgroup analysis will be provided by stratification factors and the important baseline covariates for supportive purpose. Cox regression models adjusted for randomization factors and the important baseline covariates will also be conducted as supportive analysis.

Cytogenetic subgroup analysis: It is anticipated that cytogenetic risk factors by FISH will not be determined for 10% or less of the patients (648 patients with cytogenetic results). While these patients are included in the primary analysis, PFS comparison of the two arms among patients with cytogenetic results and among those with standard risk cytogenetics are also of interest. Using the same design specifications in section 4.1.4 and assuming 648 patients have cytogenetic results, there will be at least 87% power to detect increase in median PFS among the patients with cytogenetics evaluated (296 failures, two-sided log-rank test with an overall type I error rate of 5%). Assuming that among the 648 patients with cytogenetic risk identified 70% (n=453) are standard risk, there is at least 73% power to detect a hazard ratio of 1.43 (207 failures, two-sided log-rank test with an overall type I error rate of 5%).

Program project related analyses: Samples are collected on this study and on the IFM study to evaluate correlative objectives such as defining the genetic prognostic groups evaluated by gene expression profiling and examining the best treatment in each gene expression group. These analyses are part of a program project and the detailed information on these analyses are provided in the grant and in section 14 (Other Planned studies). The DFCI and IFM data will be combined for these correlative analyses. It is expected that samples will be available for approximately 70% of the patients for correlative science data. It is not expected that treatment information will be released until after the primary final analysis. Release of data for these analyses requires DMC approval.

9. Study Subjects

9.1 Disposition of Subjects and Withdrawals

All subjects who provide informed consent will be accounted for in this study. The frequency and percent of subjects in each population, study withdrawals, subgroups and major protocol violations will also be presented. The case study report will provide the information:

- Total number of patients registered to the study
- The number not randomized and reasons why they were not randomized
- The number of patients randomized to each treatment group and whether they received assigned treatment. If they did not receive assigned treatment, the reasons why will be itemized.
- The number of patients who did not complete treatment. The number of patients who were lost to follow-up and reasons why. The number of patients who discontinued treatment and the reasons why.
- The number of patients who completed treatment in each arm.

10. Demographics and Other Baseline Characteristics

Baseline demographics and disease characteristics data including age, gender, ECOG status, disease stage, myeloma type, baseline cytogenetic data and lab measurements etc will be summarized in all subjects and by treatment groups in the ITT population. Qualitative data will be presented as frequencies and percentages. Quantitative data will be summarized as mean, standard deviation, median, interquartile range, and range.

11. Treatment

Data on the treatment administration will be summarized by treatment group and by treatment stage (RVD cycles, stem cells collection and transplant and maintenance). The actual dose, duration of treatment, relative dose intensity of each of the components of study treatment will be summarized using descriptive statistics (mean and STD, median and range). Dose modification, reason for dose modification, and reason for early treatment discontinuation will be summarized as frequency and percentage.

Concomitant medication will be summarized by treatment group similarly.

12. Efficacy Analysis

12.1 Primary Efficacy Endpoint Analysis: Progression-free survival Analysis

PFS is defined as the time from randomization to the disease progression or death from any cause. Patients who have not progressed or died are censored at the date last known progression-free. The approach that will be used to handle censoring under various scenarios is summarized in Table 1.

At the time of the interim analysis, if a case does not have final status from the response review committee the institutions assessment of response per the IMWG will be used in the analysis and will follow Table 1 for coding based on the institutions assessment.

The primary analysis will be intent to treat analysis of all randomized patients. Subjects will be analyzed according to how they are randomized regardless of the actual treatment. Cases determined to be ineligible after randomization will be included in the analysis. A sensitivity analysis of PFS will also be conducted in the ATP to confirm the robustness of the primary PFS analysis.

PFS in each of the arms will be estimated using the method of Kaplan and Meier. Median and 95% confidence interval (CI) will be provided. PFS between the two arms will be compared using two-sided stratified log-rank test. Hazard ratio as well as 95% CI for treatment will be estimated using the stratified Cox proportional hazard model with a single treatment covariate. Efron's likelihood approximation will be used to account for ties in event times and the Wald chisq test will be used to assess the treatment effect on PFS in the Cox regression. The stratification factors defined at randomization will be employed in both the stratified log-rank test and the stratified Cox model. At the final analysis, a two-sided p-value (from stratified log-rank test) which is less than 0.0455 will indicate that the two arms are different with respect to the primary endpoint of progression-free survival.

Subgroup analysis will be provided by stratification factors and the important baseline covariates for supportive purpose. Results from subgroup analysis will be presented using a forest plot. Cox

regression models adjusted for randomization factors and the important baseline covariates will also be conducted as supportive analysis.

12.2 Secondary Endpoint Analyses

The secondary objectives of the study are to compare the following outcomes between the two arms: the response rates (CR, at least VGPR), time to progression, overall survival, toxicity, quality of life and pharmaco-economics. Prognostic groups defined by gene expression profiling will be defined in correlative studies.

Response rate and duration of response: Response will be assessed based on the IMWG criteria. Response rate between treatment groups will be analyzed in the ITT population. Subjects with unknown, missing or unevaluable responses will be considered as non-responders. Subjects who discontinue from the study without prior evidence of a response to treatment will be considered as non-responders. The proportion of patients with CR, proportion of patients with at least a CR/nCR and the proportion of patients with at least a VGPR will be estimated for each treatment group with a point estimate and a 95% exact binomial confidence interval. Comparisons between the two arms will be conducted using Fisher's exact test and the exact Cochran-Mantel-Haenszel test stratified by randomization factors, with a significance level of 0.05 (two-sided). Logistic regression will be conducted to adjust for stratification factors and the important baseline characters. For response there is at least 80% power to detect differences of at least 12% between the 2 arms (Fisher's exact test, two-sided significance level of 0.05). Duration of response is defined as the time from beginning of response to progression or death. Patients who are still responding will be censored at last assessment. Duration of response will be analyzed based on data from confirmed responder (CR or at least a CR/nCR or at least VGPR) in the ITT population. The median duration of response will be estimated using the Kaplan-Meier method and compared using a stratified log-rank test.

Time to progression (TTP): TTP is defined as the time of randomization until progression. Patients who have died without evidence of progression are censored in the TTP analysis and patients who are alive without progression are censored at the last disease assessment. Details of censoring are provided in Table 1. TTP will be analyzed in the ITT population. Analysis methods will be similar to those described for the PFS analysis. A nominal p-value ≤ 0.05 (two-sided) is considered as statistical significance for TTP analysis.

Overall survival (OS): Overall survival is defined as the time of randomization to death for any cause. Patients are censored at the time last known alive. OS will be analyzed in the ITT population. Analysis methods will be similar to those described for the PFS analysis. A nominal p-value ≤ 0.05 (two-sided) is considered as statistical significance for OS analysis.

Table 1. Censoring rules for PFS, TTP, EFS

	Outcome			
Scenario	PFS	TTP	EFS	Date used in analysis
Documented disease progression	Event	Event	Event	Assessment date on which progression is first documented
Deaths[1]	Event	Censored	Event	Date of death
No disease progression and no death[2]	Censored	Censored	Censored	Last disease assessment date
Initiation of new treatment without progression or death coded[3]	Censored	Censored	Event	Date of initiation of new treatment

1. If death occurs beyond 1 year of last disease evaluation then censored at date of last disease evaluation
2. Subjects who start non-protocol therapy prior to the last disease evaluation fall under the initiation of new treatment category and are censored at the date of the initiation of new treatment or date ended treatment if non-protocol treatment date is unknown.
3. If new treatment within 1 month prior to progression, patients were coded as having an event at time of progression for PFS, TTP and EFS

QOL Endpoints : The QOL will be measured using the The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire - Core (QLQ-30), the MY20 questionnaire, and Functional Assessment of Cancer Therapy/Gynecologic Oncology Group-Neurotoxicity (FACT-NTX) side-effects questionnaires. The questionnaires will be administered on 9 occasions: baseline, time of randomization, prior to cyclophosphamide administration, 60 days post cyclophosphamide administration, 120 days post cyclophosphamide administration, at 1 year, 2 years and 3 years from baseline, and at the end of study.

The EORTC QLQ30 and MY20 will be scored according to the guidelines provided by the EORTC administration and scoring manual (EORTC Scoring Manual, 2001). For each domain and item, a linear transformation is performed to standardize the score between 0-100. Higher values indicate higher functioning and health-related quality of life. Higher values for symptoms scales indicate greater levels of symptoms. The FACT-NTX will be scored in accordance with their scoring guidelines (Cella, et al 1997). The FACT-NTX scores range 0-44, with lower values indicating higher fatigue and neurotoxicity.

For each treatment group, calculated scores and changes from baseline will be summarized by visits. Summaries of symptom-specific items of the EORTC will also be presented by treatment group and visit. Changes of scores from baseline will be compared between treatment groups using a two-sided t-test with Bonferroni justification for multiple comparisons. ANOVA model will be conducted to include the stratification factors and the important baseline characteristics in the model as appropriate. The effect of treatment will also be evaluated using a repeated measures model to incorporate all assessments across time into a single analysis. Multiple imputation methods will be used to account for missing data . However, no imputation will be implemented if answers to a whole questionnaires are missing at a certain time point.

The QOL domains to be compared between Arm A and Arm B include health related quality of life, distress, psychological functioning, physical well-being and functional well-being. Power estimates

are based on the number of complete cases; and therefore, are conservative. Differences which can be detected with 80% power between the two arms in the change of scores from time of randomization are computed using a two-sided t-test with a 0.05/8 significance level assuming 1) standard deviations of 4, 8 and 12, 2) correlations of 0.5 and 0.8 between measurements, 3) assuming that the number of patients with QOL assessments is 100, 200, 400 and 700 with equal number of patients per arm. The effect sizes which can be detected for 100, 200, 400 and 700 patients are 0.73,0.51, 0.36 and 0.25, respectively. These effect sizes translate into the following differences in QOL outcomes using the standard deviation of the scores and the correlation of repeated measures. Assuming that between 100, 200, 400 and 700 participants complete the questionnaires and standard deviation of 4, the differences which can be detected with 80% power in the change of the score between the two arms is 3, 2, 1.4 and 1.0, respectively, assuming the correlation is 0.5 and is 1.1, 0.8,0.6 and 0.4, assuming the correlation is 0.8. With a standard deviation of 8, the differences which can be detected with 80% power in the change of the score between the two arms of 5.8, 3.1, 2.9 and 2.0 , respectively, assuming correlation is 0.5 and 2.3,1.6, 1.2 and 0.8 assuming the correlation is 0.8, respectively for 100, 200, 400 and 700 participants. With a standard deviation of 16, the differences which can be detected with 80% power in the change of the score between the two arms of 8.8, 6.1, 4.3 and 3.0, respectively, assuming correlation is 0.5 and 3.5, 2.5, 1.7 and 1.2, assuming the correlation is 0.8, respectively for 100, 200, 400 and 700 participants.

Pharmacoeconomic Endpoints: The medical resource utilization will be evaluated in the two arms for U.S. study participants. Comparisons of continuous quantities will be done using the Wilcoxon-rank sum test. With an estimated 350 participants per arm, there is 80% power to detect an effect size of 0.212 using a two-sided test with a 0.05 significance level.

13. Safety and Tolerability Analyses

All adverse events recorded during study treatment will be listed by individual subjects and summarized by treatment groups as treated in the safety population. The incidence of adverse events (new or worsening from baseline) will be summarized by type of adverse event (MedDRA system organ class and preferred term), severity (based on CTC version 4 grades), attribution (relation to the study drug) by treatment group. The maximum grade will be used if multiple reports of a specific toxicity for an individual patient.

Lethal toxicities, Suspected Unexpected Serious Adverse Reaction (SUSAR), any grade 4 SAE and second malignancy will be listed by patient and tabulated by type of adverse event and treat group.

Laboratory data will be categorized and graded based on CTCAE version 4.0. The percentage of patients with laboratory abnormalities by grades will be tabulated for each treatment group. The change of grade during the study from baseline will be provided by a lab shift table, where the percentage of patients who improve or worse from baseline for each laboratory test will be

summarized for each treatment group. Descriptive statistics will also be used for those not gradable laboratory parameters.

The difference in the rate of all grade 3 or higher toxicities will be compared between the two groups using Fisher's exact test. There is at least 80% power to detect differences in the 2 arms of at least 12% in more common toxicities (>20%) and differences of at least
8% for rare toxicities (<10%), assuming two-sided Fisher's exact test, significance level of 0.05). The median time to onset of toxicities (i.e. first occurrence of a grade 3-5 AE) along with 95% CI will be estimated using Kaplan-Meier methods by treatment group. The median time to recovery (to grade 0 or 1) from the onset of toxicities along with 95%CI will be presented by treatment group.

Dose modification due to adverse experiences will be summarized**.** Data from other tests (e.g. ECG, vital signs) will be listed; notable values will be flagged and reported as appropriate.

14. Other Planned Analysis

14.1 Cytogenetic subset analysis : It is anticipated that cytogenetic risk factors by FISH will not be determined for 10% or less of the patients. While these patients are included in the primary analysis, PFS comparison of the two arms among patients with cytogenetic results and among those with standard risk cytogenetics are of interest. Stratified log-rank test will be used to compare PFS between the treatment arms for patients with cytogenetic results.

14.2 Program project related analyses:

A central theme to the program project is to evaluate the role of conventional risk factors as well as newly reported combinations of ISS and FISH abnormalities in predicting PFS and OS and whether the addition of serum free light chain, immunophenotyping and molecular techniques, or combinations thereof, to the definition of CR results in improved prediction of PFS and OS. These analyses are part of a program project and the detailed information on these analyses are provided in the grant.

For these analyses, data from the DFCI study and from the IFM study will be combined. To compare outcomes by various baseline stratification factors including ISS and FISH results, patients with baseline clinical factors will be included. For other correlative science data, it is expected that 70% of the samples from these patients will have adequate samples. Analysis from the US data alone may also be conducted.

Analysis plan will involve univariate analysis correlating factors with response and/or PFS/OS using Fisher's exact test and log-rank test, respectively. Multivariate models will include logistic (conditional and unconditional) and Cox proportional hazards models. Measures of model fit (Bayes information criteria), discrimination (c-index, reclassification measures and extension of these measures) and calibration (Hosmer-Lemeshow goodness of fit) will be used to compare the models using reported prognostic factors (ISS, FISH, gene signatures) and new ones developed from this program project. Survival regression tree methods will be used to select combinations of the factors and to identify interaction between factors. Risk prediction model characteristics will be evaluated using receiver operator characteristic methods. We will also use reclassification measures to evaluate

if the different models (simple to complex) will result in patients who are reclassified to a higher risk group "correctly" or whether the change was due to chance.

It is assumed that 1000 patients are included in the following analyses.

Differences in outcome by baseline stratification factors: For these analyses 658 failures are assumed to occur among 1000 patients. With 658 failures, there is 80% power to detect hazard ratios of 1.2 if equal allocation to 2 groups, 1.4 if 10% vs. 90% of patients are in the 2 groups and 1.7 if 5% vs. 95% are in two groups (two-sided α=0.05, logrank test). Power calculations for 3 group comparisons adjust for multiple comparisons (two-sided α=0.05/3) and assume that to compare two of the three groups 50% or 67% of the failures are in these two groups. Assuming 50% of failures are in two groups, there is 80% power to detect hazard ratios of 1.4 between two groups if equal allocation and 1.8 if 10% vs. 90% of patients are in the two groups (330 failures, log-rank test, two-sided α=0.05/3). Assuming 67% of failures are in two groups, there is 80% power to detect hazard ratios of 1.4 between two groups if equal allocation and 1.7 if 10% vs. 90% of patients are in the two groups (438 failures, log-rank test, two-sided α=0.05/3). For OS, it is anticipated that 348 deaths (assuming 3 year estimate of 75%, other assumptions same as for PFS). For 2-group comparisons, there is 80% power to detect hazard ratios of 1.4 for equal allocation, 1.6 if 10% vs 90% of patients and 2.0 if 5% vs. 95% are in the two groups (log-rank test, two-sided α=0.05).

Differences in outcome by correlative science data: For other correlative science data, it is expected that 700 samples from these patients will have adequate samples and 460 failures (deaths or progression) will occur over the 6 years of the trial (assuming 700 patients with adequate sample for correlative science data are entered over 36 months with 36 months of follow-up, exponential assumption with uniform accrual and median of 30 vs 39 months in the two arms). To evaluate the correlation of the research data with PFS, the data will be split into three groups with 460/3 failures in each group.. For 2-group comparisons, there is 80% power to detect hazard ratios of 1.6 if equal allocation and 2.1 if 10% vs. 90% of patients are in the two groups (152 failures, log-rank test, two-sided α=0.05). With 230 failures, there is 80% power to detect hazard ratios of 1.4 if equal allocation and 1.9 if 10% vs. 90% of patients are in the two groups. With 460 failures, there is 80% power to detect hazard ratios of 1.3 if equal allocation and 1.5 if 10% vs. 90% of patients are in the two groups. These are not unreasonable based on the literature where the estimated relative risks for death reported for t(4;14), del(17p) and combination of these two abnormalities with ISS were at least 2.1 and as high as 4.2 for the high risk group There would be limited power to detect risk ratios in the range of 1.7-2.5 with 152 failures if there is a risk group that is <5% of the population however, if this were to happen it is likely that the estimated relative risk would need to be sufficiently high to justify this as a separate risk group. Power calculations for 3 group comparisons adjust for multiple comparisons (two-sided α=0.05/3) and assume that to compare two of the three groups 50% or 67% of the failures are in these two groups. Assuming 50% of failures are in two groups, there is 80% power to detect hazard ratios of 2.1 between two groups if equal allocation and 3.4 if 10% vs. 90% of patients are in the two groups (76 failures, log-rank test, two-sided α=0.05/3). Assuming 67% of failures are in two groups, there is 80% power to detect hazard ratios of 1.9 between two groups if equal allocation and 2.9 if 10% vs. 90% of patients are in the two groups (101 failures, log-rank test, two-sided α=0.05/3). Similar chronological times will be used for analyses of overall survival, however, power at the initial time point will be limited. It is anticipated that there will be with 65, 138 and 243 deaths at 30, 46 and 72 months (assuming 3 year estimate of 75%, other assumptions

same as for PFS). For 2-group comparisons, there is 80% power to detect hazard ratios of 2.0-3.2, 1.6-2.2, 1.4-1.8 for 65, 138 and 243 failures for equal allocation and if 10% vs 90% of patients are in the two groups (log-rank test, two-sided α=0.05).

Response analyses using the IFM/DFCI 2009 study will proceed in a similar manner but with the 700 patients split into three equal groups (233 patients) by year of accrual with evaluation of the correlation of research data. With equal allocation of the 233 to two groups, there is at least 80% power to detect differences in overall response ranging from 15-19% depending for overall responses ranging from 30-70% (Fisher's exact test, two-sided α=0.05, constrained by requirement that response in individual groups average to the overall response). For unequal allocation, the minimal detectable difference increases as the groups become more unbalanced. For example, with 10% of the patients in one group and 90% in another the minimal detectable difference ranges from 26-32% for overall responses ranging from 30-70%; whereas with 30% in one group and 70% this difference ranges from 19-21% for overall response ranging from 30-70%.

Prediction of PFS/OS by refinement of the CR criteria

Stringent CR as well as other approaches to refine the CR using flow cytometry and molecular techniques will be evaluated in terms of differences in PFS/OS. Patients with a CR (n=300 assuming a CR rate of 30%) will be further classified as having a) normal sFLC defined as sFLC ratio of 0.26-1.65 (FLC CR, denoted as fCR), b) normal flow defined as absence of phenotypically aberrant PC in bone marrow with a minimum of 3000 total PC analyzed by multiparametric flow cytometry with U$\geq$U4 colors (immunophenotypic CR, denoted as iCR), or no minimal residual disease by molecular techniques (molecular CR, denoted as mCR). These 3 refinements of the CR category will be evaluated individually in terms of outcome (for example, compare median PFS for patients with fCR- vs fCR+) but also in combination (for example, compare PFS/OS for patients with normal SFLC ratio and normal flow vs. all others). Analysis plan is to first evaluate concordance between the four classifications (CR,fCR,iCR and mCR). The proportion of patients who are reclassified using the fCR, iCR and mCR relative to the CR will be reported with confidence intervals. Measures of model fit (Bayes information criteria), discrimination (c-index, reclassification measures and calibration (Hosmer-Lemeshow goodness of fit)) will be used to compare the models with the various classifications. Survival regression tree methods will be used to select combinations of the three refinements. The reclassification measures are of particular interest here because they will enable us to evaluate if the patients who move from CR to either a fCR, iCR, mCR (or any combination of these determined from regression tree analysis) were reclassified "correctly" or whether the change was due to chance. Over the 6 years of the RVD trial is expected that among 300 CR patients that there will be 125 PFS failures (assuming 300 patients entered over 36 months with 36 months of follow-up, exponential assumption with uniform accrual and median of 50 vs 59 months for patients with CR in the two arms). For 2-group comparisons, there is 80% power to detect hazard ratios of 1.7 if equal allocation and 2.3 if 10% vs 90% of patients are in the two groups (125 failures, log-rank test, two-sided α=0.05). If the number of CR patients is closer to 400, the number of anticipated PFS/OS failures is 167 and there is 80% power to detect hazard ratios of 1.5 if equal allocation and 2.1 if 10% vs 90% of patients are in the two groups (167 failures, log-rank test, two-sided α=0.05).

14.3. Proposed analysis for both the IFM and US Studies:

A descriptive analysis is proposed to attempt to address a question comparing PFS and OS for maintenance lenalidomide for 1 year (700 patients) vs. maintenance lenalidomide until PD (720

patients) overall and by conventional dose (350 vs. 360 patients) /high dose therapy arm (350 vs. 360 patients). This analysis is descriptive only because patients are not randomized to the 1 year vs. extended maintenance and due to the complete confounding by country (IFM vs. US) of maintenance therapy duration..

15. References

1. US Federal Register (1998) International Conference on Harmonization : Guidance on Statistical Principles for Clinical Trials. Department of Health and Human Services : Food and Drug Administration [Docket No. 97D-0174]. Federal Register Volumn 63, Number 179, pages 49583-49598. September 16, 1998.
2. ASA (1999). Ethical Guidelines for Statistical Practice. Prepared by the Committee on Professional Ethics, August 7, 1999. http://www.amstat.org/about/ethicalguidelines.cfm
3. RSS (1993). The Royal Statistical Society : Code of Conduct, August 1993. http://membership.rss.org.uk/main.asp?page=1875

16. Index of Planned Tables

Table 1.	Accrual by country
Table 2.	Accrual by country and Institution
	Accrual by year
Table 3a	Case Status on Initial Therapy (RVD cycle 1) with itemized reasons patients did not continue to randomization summarized overall and by country.
Table 3b	Case Status on Arm A (RVD) and Arm B (RVD+transplant)
Table 3c	Case Status on Maintenance Therapy for Arm A and Arm B
Table 4a:	Patient characteristics for eligible patients on Initial Therapy
Table 4b	Patient characteristics for eligible patients randomized to Arm A vs. Arm B.
Table 5a.	Toxicity for Arm A (RVD) vs. Arm B (RVD+transplant) during all treatment
Table 5b	Toxicity for Arm A (RVD) vs. Arm B(RVD+transplant) prior to maintenance
Table 5c	Toxicity for Arm A (RVD) vs. Arm B(RVD+transplant) during maintenance only
Table 5d:	Lethal Toxicities if appropriate
Table 5e	Second Malignancies if appropriate
	Any grade 3 or 4 toxicities
	Reasons off study by arm and by treatment cycle
Table 6	Primary Efficacy Endpoint. Progression-free survival (PFS) from time of randomization for all patients randomized to Arm A vs. Arm B. PFS Estimates, estimated relative risk and 95% confidence intervals are presented across all patients and by stratification factors (International Staging Criteria (stage I, II or III), cytogenetics (standard vs. high risk vs. FISH failures), country (IFM vs. US).
Table 7a:	Secondary Efficacy Endpoint: Response per the IMWG criteria for all patients randomized to Arm A vs. Arm B.
Table 7b:	Secondary Efficacy Endpoint: Response per the Blade criteria for all patients randomized to Arm A vs. Arm B.
Table 7c.	Secondary Efficacy Endpoint: Response per the sFLC criteria for all patients randomized to Arm A vs. Arm B.
Table 7d.	Secondary Efficacy Endpoint: Cross-tabulation of IMWG criteria, Blade and sFLC criteria.
Table 8	Secondary Efficacy Endpoint. Time to progression (TTP) from time of randomization for all patients randomized to Arm A vs. Arm B. TTP Estimates, estimated relative risk and 95% confidence intervals are presented across all patients and by stratification factors (International Staging Criteria (stage I, II or III), cytogenetics (standard vs. high risk vs. FISH failures), country (IFM vs. US).
Table 9	Secondary Efficacy Endpoint. Overall survival from time of randomization for all patients randomized to Arm A vs. Arm B. OS Estimates, estimated relative risk and 95% confidence intervals are presented across all patients and by stratification factors (International Staging Criteria (stage I, II or III), cytogenetics (standard vs. high risk vs. FISH failures), country (IFM vs. US)
Table 10a.	Quality of Life Endpoints-Number (%) of patients with assessments completed at each of the time points for Arm A and Arm B for the EORTC QLQ-30 and EORTC QLQ-MY20. Results summarized for both countries combined and separately by country.
Table 10b	Quality of Life Endpoints-Number (%) of patients with assessments completed at each of the time points for Arm A and Arm B for the FACT-GOG Neurotoxicity. Questionnaire. Results summarized for both countries combined and separately by country.

Table 10c	Quality of Life Endpoints-Baseline characteristics among the patients who have either C1D1/C2D1 and at least 1 follow-up EORTC questionnaire completed for Arm A vs. Arm B. Results summarized for both countries combined and separately by country.
Table 10d	Quality of Life Endpoints-C1D1 EORTC scores for Arm A and Arm B. Results summarized for both countries combined and separately by country.
Table 10e	Quality of Life Endpoints-C2D1 EORTC scores for Arm A and Arm B. Results summarized for both countries combined and separately by country.
Table 10f	Quality of Life Endpoints- Change in EORTC scores by assessment time for Arm A and Arm B. Results summarized for both countries combined and separately by country.
Table 10g	Quality of Life Endpoints-Depending on the amount of missing information and the causes for the missing information there may be additional tables included.
Table 11	Pharmacoeconomic Endpoints-
Table 12	Subset Analysis

17. Index of Planned Figures

Figure 1	Progression-free survival (PFS) from time of randomization for all patients randomized to Arm A vs. Arm B.
Figure 2a	Progression-free survival (PFS) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: International Staging Criteria (stage I, II or III),
Figure 2b	Progression-free survival (PFS) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: cytogenetics (standard vs. high risk vs. FISH failures)
Figure 2c.	Progression-free survival (PFS) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: country (IFM vs. US).
Figure 3a	Time to Progression (TTP) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: International Staging Criteria (stage I, II or III),
Figure 3b	Time to Progression (TTP) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: cytogenetics (standard vs. high risk vs. FISH failures)
Figure 3c.	Time to Progression (TTP) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: country (IFM vs. US).
Figure 4a	Overall Survival (OS) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: International Staging Criteria (stage I, II or III),
Figure 4b	Overall Survival (OS) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: cytogenetics (standard vs. high risk vs. FISH failures)
Figure 4c.	Overall Survival (OS) from time of randomization for all patients randomized to Arm A vs. Arm B by stratification factor: country (IFM vs. US).
Figure 5	Global Health Score from the QLQ-C30 over time by Arm A vs. Arm B
Figure 6	Score from the EORTC QLQ-MY20

18 Appendices

18.1 Appendix 1 Study Contact Information

Contact information for the study Principal Investigators is listed below.

Paul Richardson, MD
Dana-Farber Cancer Institute
44 Binney Street, DA1B309
Boston, MA 02115
Phone: [redacted]
Fax: [redacted]
Cell :
Email: [redacted]

Contact information for study Project Managers is listed below.

[redacted]
Dana-Farber Cancer Institute
44 Binney Street, LG-LC12
Boston, MA
Phone: [redacted]
Fax: [redacted]
Email: [redacted]

Contact information for members of the data analysis group is listed below.

[redacted]
Dana Farber Cancer Institute
44 Binney Street, CLS11007
Boston, MA 02115
Phone: [redacted]
Fax: [redacted]
Email: [redacted]

[redacted]
44 Binney Street, CLS11023
Boston, MA 02115
Phone: [redacted]
Fax: [redacted]
Email: [redacted]

18.2 Appendix 2. Multiple Myeloma Response Criteria – See the protocol section 10.

International Myeloma Working Group Response Criteria

Response criteria for all categories and subcategories of response except CR are applicable only to patients who have 'measurable' disease, as defined in Section 10.2.2.
All response categories require two consecutive assessments made at anytime before the institution of any new therapy; all categories also require no known evidence of progressive or new bone lesions if radiographic studies were performed. Radiographic studies are not required to satisfy these response requirements.

Stringent CR: CR as defined below plus normal free light chain ratio and absence of clonal cells in bone marrow* by immunohistochemistry or immunofluorescence.**

*Confirmation with repeat bone marrow biopsy is not needed.

**Presence/absence of clonal cells is based upon the k/λ ratio. An abnormal k/l ratio by immunohistochemistry and/or immunofluorescence requires a minimum of 100 plasma cells for analysis. An abnormal ratio reflecting presence of an abnormal clone is k/λ of > 4:1 or < 1:2.

CR: Negative immunofixation on the serum and urine and disappearance of any soft tissue plasmacytomas and ≤5% plasma cells in bone marrow.

*Confirmation with repeat bone marrow biopsy is not needed.

VGPR: Serum and urine M-protein detectable by immunofixation but not on electrophoresis or 90% or greater reduction in serum M-protein plus urine M-protein level <100mg per 24
hours.

PR: ≥ 50% reduction of serum M-protein and reduction in 24-h urinary M-protein by > 90% or to < 200mg per 24 hours. If the serum and urine M-protein are unmeasurable, a ≥ 50% decrease in the difference between involved and uninvolved free light chain levels is required in place of the M-protein criteria (definition of measurable disease in Section 10.2.3). If serum and urine M-protein are unmeasurable, and serum free light assay is also unmeasurable, ≥ 50% reduction in plasma cells is required in place of M-protein, provided baseline bone marrow plasma cell percentage was ≥ 30%. In addition to the above listed criteria, if present at baseline, a > 50% reduction in the size of soft tissue plasmacytomas is also required.

SD: Not meeting criteria for CR, VGPR, PR or progressive disease. This is not recommended as an indicator of response; stability of disease is best described by providing the time to progression estimates.

PD: > 25% increase of serum M-protein (which must also be an absolute increase of ≥ 0.5 g/dL) and/or urine M-protein (which must also be an absolute increase of ≥ 200 mg/24hr). If serum and urine M-protein are unmeasurable, there must be an absolute increase of ≥ 10 mg/dL between involved and uninvolved FLC levels. PD is also measured by an absolute increase in bone marrow plasma cells ≥ 10%. In addition to the above listed criteria, progression may also be measured by a definite development of new bone lesions or soft

tissue plasmacytomas or definite increase in the size of existing bone lesions or soft tissue plasmacytomas or development of hypercalcemia (corrected serum calcium $\geq$ 11.5 mg/dL or 2.65 mmol/L) that can be attributed solely to the plasma cell proliferative disorder.

10.2.4.2 FreeLite™ Disease Response Criteria

Complete Response: For those patients being followed by serum free light chain (and NO measurable serum or urine M-spike), which were immunofixation negative at enrollment, normalization of serum free light chain ratio.

- Normalization is defined as the serum free light chain ratio being within the normal range. If the serum free light chain ratio is not within the normal range, but the individual kappa and lambda light chain values are within normal range, this may be considered CR.

Partial Response: If only measurable parameter is serum immunoglobulins free light chain (FLC), EITHER of the following changes qualify as partial response:

- A 50% decrease in the difference between involved and uninvolved FLC levels; OR

- A 50% decrease in the level of involved FLC AND a 50% decrease (or normalization) in the ratio of involved/uninvolved FLC

Minimal Response: 25 – 49% reduction in the level of the serum monoclonal paraprotein. Patients being followed by serum immunoglobulins free light chain only will not be assessed for MR category.

Progressive Disease: If only measurable parameter is serum immunoglobulins free light (FLC), either of the following qualify as progression:

- 50% increase in the difference between involved and uninvolved FLC levels from the lowest response level, which must also be an absolute increase of at least 10 mg/dL; OR

- 50% increase in the level of involved FLC AND a 50% increase in the ratio of involved/uninvolved FLC from the lowest response level.

Modified EBMT Response Criteria	
Response	**Criteria for Response[a]**
Complete response (CR)	Requires all of the following: Disappearance of the original monoclonal protein from the blood and urine on at least two determinations for a minimum of six weeks by immunofixation studies. <5% plasma cells in the bone marrow on at least two determinations for a minimum of six weeks. [b] No increase in the size or number of lytic bone lesions (development of a compression fracture does not exclude response).[c] Disappearance of soft tissue plasmacytomas for at least six weeks.
Near Complete Response (nCR)	Requires the following: Same as CR, but immunofixation studies continue to show presence of the monoclonal protein
Very Good Partial Response (VGPR)	Requires the following: ≥ 90% reduction in serum M-protein plus urine M-protein level <100mg per 24 hours on at least two determinations for a minimum of six weeks.
Partial response (PR)	PR includes participants in whom some, but not all, criteria for CR are fulfilled providing the remaining criteria satisfy the requirements for PR. Required all of the following: ≥50% reduction in the level of serum monoclonal protein for at least two determinations six weeks apart. If present, reduction in 24-hour urinary light chain excretion by either ≥90% or to <200 mg for at least two determinations six weeks apart. ≥50% reduction in the size of soft tissue plasmacytomas (by clinical or radiographic examination) for at least six weeks. No increase in size or number of lytic bone lesions (development of compression fracture does not exclude response).[c]
Minimal response (MR)	MR included participants in whom some, but not all, criteria for PR were fulfilled, providing the remaining criteria satisfied the requirements for MR. Required all of the following: ≥25% to ≤ 49% reduction in the level of serum monoclonal protein for at least two determinations six weeks apart. If present, a 50 to 89% reduction in 24-hour light chain excretion, which still exceeds 200 mg/24 h, for at least two determinations six weeks apart. 25-49% reduction in the size of plasmacytomas (by clinical or radiographic examination) for at least six weeks. No increase in size or number of lytic bone lesions (development of compression fracture does not exclude response).[c]
No change (NC)	Not meeting the criteria for MR or PD.

Modified EBMT Response Criteria	
Response	**Criteria for Response[a]**
Progressive disease (PD) (for participants not in CR)	Requires one or more of the following: >25% increase[d] in the level of serum monoclonal paraprotein, which must also be an absolute increase of at least 5 g/L and confirmed on a repeat investigation. >25% increase[d] in 24-hour urinary light chain excretion, which must also be an absolute increase of at least 200 mg/24 h and confirmed on a repeat investigation. >25% increase[d] in plasma cells in a bone marrow aspirate or on trephine biopsy, which must also be an absolute increase of at least 10%. Definite increase in the size of existing lytic bone lesions or soft tissue plasmacytomas. Development of new bone lesions or soft tissue plasmacytomas (not including compression fracture). Development of hypercalcemia (corrected serum calcium >11.5 mg/dL or 2.8 mmol/L not attributable to any other cause).
Relapse from CR	Required at least one of the following: Reappearance of serum or urinary paraprotein on immunofixation or routine electrophoresis confirmed by at least one follow-up and excluding oligoclonal immune reconstitution. ≥5% plasma cells in the bone marrow aspirate or biopsy. Development of new lytic bone lesions or soft tissue plasmacytomas or definite increase in the size of residual bone lesions (not including compression fracture). Development of hypercalcemia (corrected serum calcium >11.5 mg/dL or 2.8 mmol/L not attributable to any other cause)[e].

a Based on the criteria reported by Blade et al., 1998.

b Per Blade *et al*., 1998, if absence of the monoclonal protein is sustained for 6 weeks it is not necessary to repeat the bone marrow except in participants with nonsecretory myeloma where the marrow examination must be repeated after an interval of at least 6 weeks to confirm CR.

c Per Blade *et al*., 1998, skeletal X-Rays are not required for the definition of response, but if performed there must be no evidence of progression of bone disease (no increase in size or number of lytic bone lesions).

d It is suggested that the reference point for calculating any increase should be the lowest value of the preceding confirmed response (MR, PR or CR) or the baseline value if there is no previous confirmed response.

e Other clinical data may be requested by the IRC, as necessary, to assess the cause of the hypercalcemia