BeiGene

NCT03053440

STATISTICAL ANALYSIS PLAN

Study Protocol Number:	**BGB-3111-302**
Study Protocol Title:	**A Phase 3, Randomized, Open-Label, Multicenter Study Comparing the Efficacy and Safety of the Bruton's Tyrosine Kinase (BTK) Inhibitors BGB-3111 and Ibrutinib in Subjects with Waldenström's Macroglobulinemia (WM)**
Date:	Oct 18, 2019
Version:	1.0

SIGNATURE PAGE

[redacted]

Senior Manager, Biostatistics

DocuSigned by: [redacted]

Signature: [redacted] | I approve this document 18-Oct-2019 | 18:23:43 PDT

Date: 18-Oct-2019 | 18:23:43 PDT

4A3D5ADADBD14D07BAE6203BDA33719C

Approval

[redacted]

Vice President, Biometrics

DocuSigned by: [redacted]

Signature: [redacted] | I approve this document 20-Oct-2019 | 19:46:49 PDT

Date: 20-Oct-2019 | 19:46:52 PDT

C4979E19F5584C92B3FD70F23DE589AA

[redacted]

Chief Medical Officer

DocuSigned by: [redacted]

Signature: [redacted] | I approve this document 21-Oct-2019 | 14:35:13 PDT

Date: 21-Oct-2019 | 14:35:16 PDT

F0ACB807EC9D434E967D19BCBEAA7AAA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

Abbreviation	Term
AEs	Adverse events
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ATC	Anatomical therapeutic chemical
AUC	Area under the plasma concentration-time curve
BID	Twice a day
BMI	Body mass index
BOR	Best overall response
BTK	Bruton tyrosine kinase
C_{max}	Maximum observed plasma concentration
CR	Complete response
CSR	Clinical study report
CT	Computed tomography
DBP	Diastolic blood pressure
DLBCL	Diffuse large B-cell lymphoma
DOR	Duration of response

eCRF	Electronic case report form
ECG	Electrocardiogram
ECOG	Eastern cooperative oncology group
EDC	Electronic data capture
INR	International normalized ratio
IPSS	International prognostic scoring system
IRC	Independent review charter
LDT	Lab developed test
MedDRA	Medical Dictionary for Regulatory Activities
MRD	Minimal residual disease
MRR	Major response rate
NGS	Next generation sequencing
NCI-CTCAE	National Cancer Institute Common Toxicity Criteria for Adverse Events
ORR	Overall response rate
OS	Overall survival
PBMCs	Peripheral blood mononuclear cells
PD	Disease progression
PFS	Progression-free survival

PK	Pharmacokinetic
PT	Preferred term
RT	Richter's transformation
Q1, Q3	First quartile, third quartile
QD	Once daily
QT	Electrocardiographic interval
SAEs	Serious adverse events
SAP	Statistical analysis plan
SBP	Systolic blood pressure
SOC	System organ class
SPD	Sum of products of diameters
SD	Standard deviation
$t_{1/2}$	Terminal half-life
TEAE	Treatment-emergent adverse event
t_{max}	Time to maximum observed plasma concentration
TTR	Time to response
VGPR	Very good partial response
WHO-DD	World Health Organization Drug Dictionary

WM	Waldenström's macroglobulinemia

1 INTRODUCTION

This statistical analysis plan (SAP) describes the detailed plans for the analysis of safety and efficacy data for study BGB-3111-302. This document is based on the protocol Version 4 dated 21-SEP-2018. The analysis plan for the pharmacogenomics and exploratory [redacted] analyses are not included in this SAP.

2 STUDY OVERVIEW

This is a Phase 3, randomized, open-label, multicenter study comparing the efficacy and safety of the Bruton's Tyrosine Kinase (BTK) Inhibitors zanubrutinib and ibrutinib in subjects with Waldenström's Macroglobulinemia (WM) who require therapy according to the consensus panel criteria from the Seventh International Workshop on Waldenström's macroglobulinemia (Dimopoulos et al 2014). Subjects may be treatment-naïve (TN) or have relapsed or been refractory to prior therapy. Relapsed means a subject previously achieved a CR or VGPR/PR but after a period of 6 months or more showed progressive disease. Refractory means a subject experienced prior treatment failure or disease progression within 6 months of therapy initiation. Subjects with no prior therapy (TN) will comprise no more than 20% of the study population in Cohort 1. The study is composed of an initial Screening Phase (up to 35 days), a Treatment Phase, and a Follow-up Phase. The study schema is presented in Figure 1.

Approximately 210 subjects will be enrolled on the study. All subjects enrolled on the study will have the *MYD88* gene sequenced by a central laboratory using a LDT assay using the baseline bone marrow. Approximately one hundred and eighty-eight WM subjects (150 relapsed/refractory subjects and approximately 38 treatment-naïve subjects) who have the $MYD88^{MUT}$ mutation, which is characteristic of WM and present in approximately 90% of cases, will be enrolled on to Cohort 1 and randomized to one of two treatment arms (Cohort 1; zanubrutinib treatment [Arm A] or ibrutinib treatment [Arm B]) in a 1:1 ratio using *CXCR4* mutational status ($CXCR4^{WHIM}$ vs. $CXCR4^{WT}$ vs missing) and number of prior lines of therapy for WM (0 vs. 1-3 vs. >3 prior therapies) as stratification factors. Subjects found to have $MYD88^{WT}$ by gene sequencing, which is estimated to be present in approximately 10% of enrolled subjects, will be enrolled to Cohort 2 and will receive zanubrutinib treatment on a third, non-randomized study arm (Arm C). This non-randomized arm is aimed to evaluate the efficacy of zanubrutinib in $MYD88^{WT}$ WM subjects, among whom suboptimal efficacy has been observed

(i.e., shorter median survival and lower MRR and CR/VGPR rates versus *MYD88*^{MUT}) when treated with ibrutinib. In addition, those subjects whose *MYD88* mutational status is missing or inconclusive will be assigned to Cohort 2, Arm C. Arm C will enroll approximately 22 subjects.

The primary efficacy analysis will be conducted at least 15 months after 90% enrollment in the RR Analysis Set was completed. At the time, 90% of subjects enrolled in Cohort 1 are expected to have at least 19 months of study follow up. Tumor response will be assessed every cycle (every 4 weeks) for the first 48 weeks and then every 3 cycles thereafter by an independent review committee (IRC) according to an adaptation of the response criteria updated at the Sixth IWWM (Owen et al 2013; NCCN Guidelines, Lymphoplasmacytic Lymphoma/Waldenström's Macroglobulinemia 2015: v2). Serum IgM and M-protein levels will be measured at Screening, on Day 1 of every cycle for the first 48 weeks (12 cycles) then every 3 cycles thereafter. For subjects with evidence of extramedullary disease (lymphadenopathy and/or splenomegaly) by computed tomography (CT) scan at baseline, assessment by CT scan will occur every 12 weeks (starting from C4D1) during the first 48 weeks (until C13D1), then every 24 weeks until PD or complete resolution of extramedullary disease. A CT scan is also required at the safety follow-up visit. Bone marrow will be assessed by aspirate and biopsy at Screening, after 12 cycles (C13D1), at time of suspected CR, and as clinically indicated.

Response assessments will occur in all subjects at the beginning of every cycle (every 4 weeks) starting from C2D1 during the first 12 cycles (48 weeks), and then every 3 cycles (every 12 weeks) thereafter e.g., C2D1, C3D1, etc. For those with extramedullary disease at baseline response assessments must occur in conjunction with CT scans until resolution of extramedullary disease. For response assessments that occur during cycles where a CT scan is not required then results from prior scans (up to 12 weeks/3 cycles during the first 48 weeks/12 cycles and up to 24 weeks/6 cycles thereafter) can be carried forward in those subjects with extramedullary disease at baseline. Quality of life (QOL) will be measured every 12 weeks during the first 48 weeks, and then every 24 weeks thereafter by the EORTC QLQ-C30 and EQ-5D in *MYD88*^{MUT} WM subjects (Cohort 1).

All subjects will be followed for AEs for 30 additional days after the last dose of study drug. All treatment-related AEs and serious AEs (SAEs) will be followed until resolution or stabilization. Efficacy evaluations will continue until documented PD for subjects who discontinued for reasons other than PD.

Screening Phase: Screening evaluations will be performed within 35 days prior to the randomization, with the exception of a fresh bone marrow biopsy, which may be performed up to 42 days prior to randomization, as long as no intervening therapy has been administered. A fresh bone marrow aspirate is required for flow cytometry and the *MYD88* and *CXCR4* mutational analyses at Screening. If subject enrolls based on a bone marrow biopsy that was obtained within the 42 days of randomization, then a fresh bone marrow aspirate will still be required during the Screening period. Subjects who agree to participate will sign the informed consent form prior to any screening evaluations. Screening procedures are outlined in the Study Assessments and Procedures Schedule in the protocol. Screening evaluations can be repeated within the screening period.

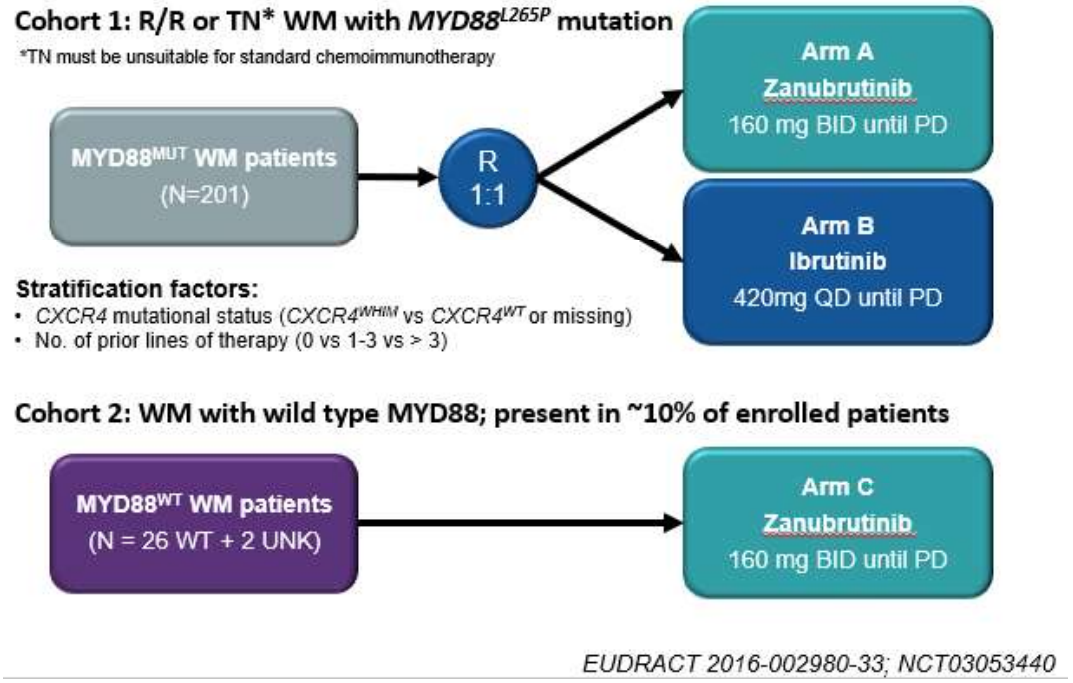
Treatment Phase: Subjects will be assigned to either Cohort 1 ($MYD88^{MUT}$) or Cohort 2 ($MYD88^{WT/\text{missing}}$) based on the mutational status of the *MYD88* gene. Cohort 1 subjects will be randomized using an Interactive Response Technology (IRT) system by the status of stratification factors to receive the first dose of zanubrutinib or ibrutinib at Cycle 1 Day 1. Cohort 2 subjects ($MYD88^{WT/missing}$) will be assigned by the IRT to receive zanubrutinib. Subjects randomized to Arm A will take 160 mg (80 mg x 2 capsules) of zanubrutinib PO BID. Subjects randomized to Arm B will take 420 mg (140 mg x 3 capsules) of ibrutinib PO QD. Subjects assigned to Cohort 2 (Arm C) will take zanubrutinib 160 mg (80 mg x 2 capsules) PO BID. All subjects will continue to be treated and followed by the defined schedules in the protocol for IgM, radiologic assessments and bone marrow, etc. until PD, unacceptable toxicity, death, withdrawal of consent, lost to follow-up, end of treatment duration or the study is terminated by the sponsor, whichever comes first. A treatment cycle consists of 28 days.

Follow-up Phase: In all study arms, subjects will return approximately 30 days after the last dose of study drug for Safety Follow-up Visit(s) for the collection of AEs and SAEs that may have occurred after the subject discontinued from the study. For subjects that had extramedullary disease at baseline, a CT scan is also required at the Safety Follow-up Visit. All treatment-related SAEs will be followed until resolution or stabilization. The investigator or his/her designee will also continue to collect information on new anticancer therapy given after the last dose of study drug. Efficacy evaluations will continue until documented PD. If study drug is discontinued due to reasons other than PD, serum IgM and M-protein levels will continue to be followed every 12 weeks (±14 days) while other efficacy evaluations will be followed as per investigator's discretion. Efficacy assessments will be continued, as per protocol, until PD,

withdrawal of consent, death, lost to follow-up, end of study or study termination by sponsor, whichever occurs first. Follow-up will continue to occur even though a subject may have started a new anticancer therapy after the last dose of study drug. For full efficacy assessment schedules, please refer to Section 7.3 and Table 2 of the protocol. Subjects will be followed for survival and further anticancer therapy information after progression of disease via phone contact (with the subject's guardian, if applicable) every 12 weeks (±14 days) until study end.

Post Study: Subjects assigned to Arms A and C (zanubrutinib) who, in the opinion of the investigator, continue to benefit from zanubrutinib or ibrutinib at study closure may continue treatment with zanubrutinib by enrolling on the zanubrutinib Long Term Extension Study. This study is a rollover study for subjects who wish to continue receiving zanubrutinib.

Figure 1. Schema for Study BGB-3111-302



3 STUDY OBJECTIVES

3.1 PRIMARY OBJECTIVE (COHORT 1)

- To compare the efficacy in terms of VGPR/CR rate of zanubrutinib vs ibrutinib in subjects with *MYD88*^{MUT} WM

3.2 SECONDARY OBJECTIVES (COHORT 1)

- To compare the efficacy in terms of major response rate (MRR) of zanubrutinib vs ibrutinib in subjects with *MYD88*^{MUT} WM
- To evaluate the efficacy in terms of duration of response and progression-free survival of zanubrutinib vs ibrutinib in subjects with *MYD88*^{MUT} WM
- To further evaluate clinical benefit, and anti-lymphoma effects of zanubrutinib vs ibrutinib in subjects with *MYD88*^{MUT} WM
- To evaluate safety and tolerability of zanubrutinib versus ibrutinib in subjects with *MYD88*^{MUT} WM, as measured by the incidence and severity of adverse events according to the National Cancer Institute (NCI) Common Terminology for Adverse Events (CTCAE) v4.03.

3.3 EXPLORATORY OBJECTIVES

4 STUDY ENDPOINTS

4.1 PRIMARY ENDPOINT

The primary endpoint is the proportion of subjects in each arm of Cohort 1 achieving either CR or VGPR, as determined by independent review committee (IRC) using an adaptation of the response criteria updated at the Sixth IWWM (Owen et al. 2013 and NCCN Guidelines, Lymphoplasmacytic Lymphoma/Waldenström's Macroglobulinemia 2015: v2).

4.2 SECONDARY ENDPOINTS

Efficacy (Cohort 1):

- Major response rate (MRR) as assessed by the IRC, defined as the proportion of subjects achieving CR, VGPR, or partial response (PR). MRR by IRC is the key secondary endpoint.
- Duration of response (DOR) as assessed by the IRC, defined as the time from first determination of response (CR, VGPR or PR) (per modified IWWM criteria) until first documentation of progression (per modified IWWM criteria) or death, whichever comes first
- Rate of CR or VGPR as assessed by the Investigator
- DOR as assessed by the Investigator, defined as the time from first determination of response (CR, VGPR or of PR) (per modified IWWM criteria) until first documentation of progression (per modified IWWM criteria) or death, whichever comes first
- Progression-free survival (PFS) as assessed by the IRC, defined as time from randomization to the first documentation of progression (per modified IWWM criteria) or death, whichever occurs first
- PFS as assessed by the Investigator, defined as time from randomization to the first documentation of progression (per modified IWWM criteria) or death, whichever occurs first

- Resolution of treatment-precipitating symptoms, defined as the absence of the symptoms that triggered initiation of study treatment (per the IWWM treatment guidelines) at any point during study treatment
- Anti-lymphoma effect, defined as any reduction in bone marrow involvement by lymphoplasmacytoid lymphocytes and/or size of lymphadenopathy and/or splenomegaly by CT scan, at any time during the course of study treatment

Safety (Cohort 1)

- The incidence, timing, and severity of treatment-emergent AEs (TEAE) according to NCI-CTCAE v.03

4.3 EXPLORATORY ENDPOINTS

MYD88^{MUT} WM
(Cohort 1)

- Change in QOL as assessed by EORTC QLQ-C30 and EQ-5D in subjects with *MYD88*^{MUT} WM (Cohort 1)
- Medical resource utilization as assessed by the number of hospitalizations, length of hospital stay, and supportive care in subjects with *MYD88*^{MUT} WM (Cohort 1)
- [redacted]

5 SAMPLE SIZE CONSIDERATIONS

The sample size calculation is based on the comparison of the primary endpoint of CR or VGPR rate in the Relapsed/Refractory analysis set in Cohort 1. Assuming RR_A=0.35 and RR_B=0.15, where RR_A and RR_B denote the CR or VGPR rate in arm A and arm B, seventy-five subjects per arm (150 total) provides a power of 0.814 in testing RR_A versus RR_B in the Relapsed/Refractory analysis set in Cohort 1 using a normal approximation to binomial test with a two-sided significance of 0.05. Assuming MRR_A=0.90 and MRR_B=0.80, the power of demonstrating non-inferiority of zanubrutinib in the Relapsed/Refractory analysis set is 85.5% when a NI margin of 8% is used.

In addition to the 150 relapsed/refractory subjects, approximately 20% (38) treatment-naïve subjects with *MYD88*^{MUT} will be enrolled in Cohort 1.

Assuming *MYD88*^{MUT} mutation is present in 90% of the enrolled subjects, a total of approximately 210 subjects will be enrolled in Cohort 1 and Cohort 2 combined.

6 STATISTICAL METHODS

6.1 ANALYSIS SETS

Analysis Sets for Cohort 1:

The Intent to Treat (ITT) Analysis Set includes all randomized subjects who are assigned to a treatment arm in Cohort 1.

The Relapsed/Refractory (RR) Analysis Set (a subset of the ITT analysis set) includes all randomized subjects with at least 1 prior line of therapy as determined by the IRT system. The Relapsed/Refractory Analysis Set will be the primary analysis set used for efficacy analyses.

The Per-Protocol (PP) Analysis Set includes subjects in the ITT Analysis Set meeting following criteria (ICH E9):

- Received any dose of randomized treatment regimen
- Had a valid post-baseline measurement for either IgM (central or local) or M-protein by SPEP (central or local)
- Didn't have any important protocol deviation.

Categories of important protocol deviation are defined in Section 6.3.2. Similarly, The Per-Protocol Relapsed/Refractory Analysis Set includes subjects in the Relapsed/Refractory Analysis Set who met the above criteria.

Criteria for exclusion from the PP Analysis Set will be determined and documented before the database lock for the primary analysis. Non-inferiority test for the secondary endpoint of MRR by IRC will be also tested in the Per-Protocol Relapsed/Refractory Analysis set and in the Per-Protocol Analysis set. .

Analysis Set for Cohort 2:

The efficacy analysis set in Cohort 2 includes all subjects who received any dose of zanubrutinib and are centrally confirmed to have $MYD88^{WT}$.

Analysis Set for Cohort 1 and Cohort 2:

The Safety Analysis Set includes all subjects who received any dose of zanubrutinib or ibrutinib. The Safety Analysis set will be used for all safety analyses.

The PK Analysis Set includes all subjects who have at least one post-dose zanubrutinib concentration.

The number and percentage of subjects in each analysis set will be summarized.

6.2 DATA ANALYSIS GENERAL CONSIDERATIONS

Descriptive statistics include n, mean, standard deviation, median, first quartile (Q1), third quartile (Q3), minimum, and maximum for continuous variables and n (%) for categorical variables.

All calculations and analyses will be conducted using SAS version 9.2 or higher.

6.2.1 Definitions and Computations

Study treatment (study drug): Study drug for this study is zanubrutinib or ibrutinib.

Study day: Study day will be calculated relative to the date of the first dose of study treatment (study day 1). For subjects not dosed, the enrollment (Arm C) or randomization date (Arm A/B) will be used instead of the first dose date. For assessments conducted on or after study day 1, study day will be calculated as (assessment date – date of study day 1 + 1). For assessments conducted before the date of study day 1, study day is calculated as (assessment date – date of study day 1). There is no study day 0.

In the situation where the event date is partial or missing, the date will appear partial or missing in the listings; Study day and any corresponding durations will be presented based on the imputations specified in Analysis Details Specification document.

Treatment duration: The treatment duration will be calculated as (date of last dose of study treatment – date of first dose of study treatment + 1).

Baseline: Unless otherwise specified, a baseline value is defined as the last non-missing value collected before the first dose of study treatment.

The last serum IgM value (or SPEP M-protein) on/before cycle 1 day 1 (C1D1) will serve as the baseline (last value prior to drug administration) for all assessments throughout the study except for subjects who have undergone plasmapheresis before cycle 1 day 1. For subjects who have undergone plasmapheresis, the last pre-plasmapheresis serum IgM/SPEP M-protein value (if available at screening) should be used. In the case of multiple prior consecutive plasmapheresis procedures during screening, the highest pre-dose value (least confounded by plasmapheresis) should be used as the baseline.

6.2.2 Handling of Missing Data

Missing data will not be imputed unless otherwise specified. Missing dates or partially missing dates will be imputed conservatively for adverse events and prior/concomitant medications/procedures as provided in Appendix A: Imputation of Missing or Partially Missing Dates.

When summarizing categorical variables, subjects with missing data are generally included in the denominator to calculate percentages unless otherwise specified. When needed, the category of "Missing" is created and the number of subjects with missing data is presented.

When summarizing continuous variables, subjects with missing data are not included in calculations unless otherwise specified.

No imputation of AE grades will be performed. TEAEs with missing CTCAE grade will only be summarized in the all-grades column.

If the assessment of the relationship of an AE to study treatments is missing, then the AE is assumed to be related to the study treatment in the safety analysis summary. No imputation will be done in the AE listings.

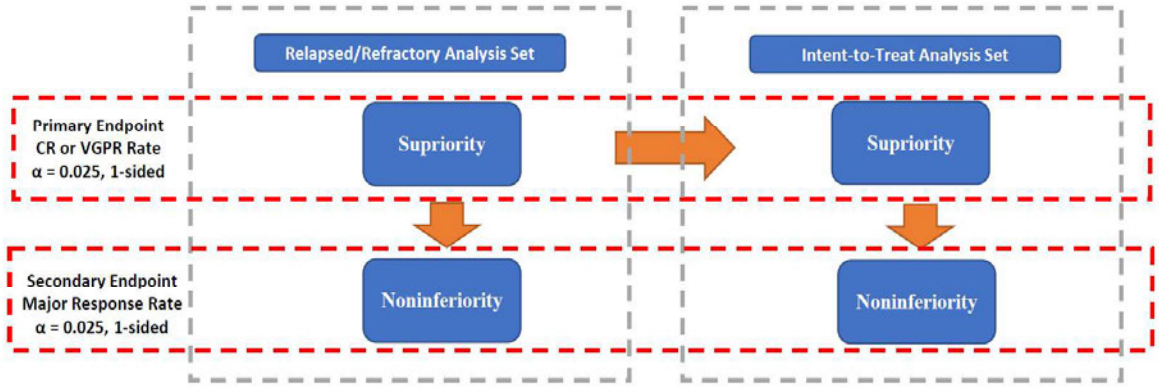
6.2.3 Adjustment for Covariates

Stratified analysis will be performed to adjust for important baseline covariates for the primary and some secondary endpoints. Details of the stratified analyses are provided in Section 6.4.

6.2.4 Multiplicity Adjustment

Multiplicity for study-wide type-I error will be adjusted for the tests of primary VGPR/CR rate endpoint and key secondary MRR endpoint in R/R analysis set and ITT analysis set. A flowchart of the multiplicity adjustment is shown in Figure 2.

Figure 2. Flowchart for the Multiplicity Adjustment



There are 2 primary hypotheses with respect to the primary endpoint of CR or VGPR rate, which will be tested by the hierarchical fixed-sequence procedure in the order of : superiority in the RR Analysis Set; and superiority in the ITT Analysis Set. These 2 hypotheses form the family of the primary hypotheses. Since each hypothesis within the family of the primary hypotheses is tested at a 1-sided 0.025 level, based on the closed test principal, the family-wise error rate (FWER) within the primary hypotheses family is controlled at a 0.025 level (1-sided).

The secondary endpoints will only be tested if the first primary hypothesis (superiority of the CR or VGPR rate in the RR Analysis Set) is significant. For the secondary endpoints, only the key secondary endpoint of MRR will be considered in controlling the study-wide FWER. The testing in the other secondary endpoints will be descriptive only. The key secondary endpoint of MRR will be tested for noninferiority in the RR Analysis Set and the ITT Analysis Set. These 2 hypotheses form the family of the secondary hypotheses and can be considered as the descents of the superiority testing of the CR or VGPR rate in the corresponding analysis sets (Bretz et al, 2009). The noninferiority in MRR in the RR Analysis Set will be tested at a 1-sided 0.025 level if the superiority of the CR or VGPR rate is demonstrated in the RR Analysis Set under this approach. If the superiority of the CR or VGPR rate can be demonstrated in both of the RR and the ITT Analysis Sets, the noninferiority of MMR will be tested in the RR and the ITT Analysis Sets respectively at a 1-sided level of 0.025.

Based on Bretz (2009), the testing strategy described above is a consistent procedure that avoids illogical dependencies. Under the above testing strategy, the family-wide error rate in both of

the primary and the secondary families of hypotheses is strongly controlled at a 1-sided level of 0.05 while the family-wide error in the primary family of hypotheses is strongly controlled at 1-sided level of 0.025 (Bretz 2009).

6.2.5 Data Integrity

Before pre-specified statistical analysis begins, the integrity of the data should be reviewed to assure fit-for-purpose. The data set for analysis should be an accurate and complete representation of the subjects' relevant outcomes from the clinical database. All essential data should be complete and reviewed up to a pre-specified cutoff date. Critical consistency checks and appropriate source data verification should be completed according to the final data extraction plan.

6.3 SUBJECT CHARACTERISTICS

6.3.1 Subject Disposition

The following subject disposition information will be summarized by arm:

- Number of subjects randomized (or number of subjects enrolled for Arm C)
- Number (%) of subjects treated
- Number (%) of subjects randomized, but not treated
    - Reason for not treated
- Number (%) of treated subjects who discontinued treatment
    - Reason for treatment discontinuation
- Number (%) of treated subjects who discontinued study
    - Reason for study discontinuation

The number of subjects randomized or enrolled will be summarized by country/site.

The number and percentage of subjects still receiving treatment and still in the study at data cutoff will also be summarized.

Study follow-up time is defined as the time from the randomization (or enrollment date for Arm C) to the death date or the end of study date for the subjects who discontinued from study

(whichever occurs first), or the data cutoff date for the ongoing subjects. Study follow-up time will be summarized descriptively (e.g. median, min, max, etc.).

6.3.2 Protocol Deviations

Important protocol deviation criteria will be established before the database lock. The following categories will be considered when evaluating a deviation as an important protocol deviation:

- Enrollment of a patient into the study even though that patient did not meet all eligibility criteria according to the current protocol at the time of the enrollment
- Failure to withdraw a patient from study treatment when that patient met protocol criteria requiring withdrawal from study treatment
- Administration of the incorrect study treatment to a study patient
- Administration of a medication considered prohibited according to the current protocol at the time of the medication administration
- Assignment of patient to the incorrect study cohort based on the study selection assay
- Prior to obtaining informed consent, performance of a study-specific procedure that is not considered a standard or typical procedure for the disease under study.

Important protocol deviations will be summarized by deviation category and by arm. A listing will also be provided.

6.3.3 Randomization Stratification Factors

In Cohort 1, the number of subjects with each of the IRT randomization stratification factors will be summarized by arm. The randomization stratification factors include the CXCR4 mutation status and the prior lines of therapy. In addition, stratification errors (defined as IRT-based strata vs data-driven strata inconsistencies) will be summarized. All the stratified analysis will be based on the strata used for randomization.

6.3.4 Demographics and Other Baseline Characteristics

Demographic and baseline characteristics including the following will be summarized by arm using descriptive statistics:

- Age (years);

- Age group (<=65 vs. >65 years);
- Gender;
- Race and ethnicity;
- Geographic region;
- Height (cm), weight (kg), and body mass index (BMI, kg/m^2);
- Vital signs;
- Eastern Cooperative Oncology Group (ECOG) performance status;
- ECG parameters (PR internal, QT interval, QTcF interval).

6.3.5 Disease History and Baseline Disease Characteristics

Baseline disease characteristics including the following will be summarized by arm using descriptive statistics:

- Time since initial WM diagnosis;
- Prognostic group at WM diagnosis and at study entry;
- Prior line of therapy for WM (0 vs. 1-3 vs. >3);
- Time since the most recent progression;
- Serum immunoglobulin (IgM, IgA and IgG);
- Serum IgM (>=40 g/L vs <40 g/L)
- SPEP M-protein;
- Serum immunofixation (positive vs. negative);
- β2 microglobulin (quantitative and qualitative: ≤3 mg/L vs. >3 mg/L);
- Genotype by LDT/Sanger method: (MYD88-MUT/CXCR4-WT vs MYD88-MUT/CXCR4-WHIM vs MYD88-WT/CXCR4-WT vs missing);
- Presence of extramedullary disease per investigator (yes: splenomegaly/lymphadenopathy vs. no);
- Presence of extramedullary disease per IRC (yes: splenomegaly/lymphadenopathy vs. no);
- Baseline bone marrow involvement %;

- WM IPSS (as derived by sponsor based on Appendix B: low, intermediate, high);
- Hemoglobin (quantitative and qualitative: ≤110 g/L vs. >110 g/L);
- Platelet (quantitative and qualitative: ≤100 10^9/L vs. >100 10^9/L);
- ANC (quantitative and qualitative: ≤1.5 10^9/L vs. >1.5 10^9/L);
- Any cytopenia (yes: hemoglobin ≤110 g/L or platelet count ≤100 10^9/L or ANC ≤1.5 10^9/L vs. no);
- Hematocrit (%);
- Prior radiotherapy and prior transplants (yes vs. no);
- WM signs and symptoms.
- Cryoglobulinemia based on central lab (positive vs negative)

Categories of baseline characteristics can be modified depending the availability of data and/or percentage of categories.

6.3.6 Medical History

Medical history will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) (version 20.0 or higher). The number and percentage of subjects reporting a history of any medical condition, as recorded on the CRF, will be summarized by system organ class (SOC) preferred term (PT), and arm. A listing of medical history will be provided.

6.3.7 Prior Systemic Anti-Cancer Therapies

Number of prior therapies, duration of last therapy, best response of last therapy, and time since the end of last therapy will be summarized for prior anti-cancer therapy by arm using descriptive statistics. A listing will be provided as well.

The therapies with the same sequence/regimen number are counted as one prior therapy.

6.3.8 Prior and Concomitant Medications

Prior and concomitant medications will be coded using the World Health Organization Drug Dictionary (WHO DD) drug codes version March 2017 or later and will be further classified to the appropriate Anatomical Therapeutic Chemical (ATC) code.

Prior medications are defined as medications that started before the first dose date. Concomitant medications are defined as medications that (1) started before the first dose of the study treatment and were continuing at the time of the first dose of the study treatment, or (2) started on or after the first dose date of the study treatment up to 30 days after the last dose date or the initiation of a new anti-cancer therapy.

The number and percentage of subjects reporting prior medications and concomitant medications will be summarized by ATC medication class Level 2, WHO DD preferred name, and arm, respectively. A listing of the prior and concomitant medications will be provided.

6.4 EFFICACY ANALYSES

Statistical testing will be performed to compare the efficacy of zanubrutinib (Arm A) and ibrutinib (Arm B) in Cohort 1. For all efficacy analyses, subjects will be analyzed according to the treatment arm to which they were randomized. Descriptive statistics will be used to report the efficacy of zanubrutinib in Cohort 2 (Arm C). Subjects with missing response assessments will be considered as non-responders and will be included in the denominator when calculating the CR or VGPR rate, MRR, and ORR. The non-responders will be excluded in the corresponding duration of response analyses.

Statistical analyses will be performed after the database is locked.

6.4.1 Primary Efficacy Analyses

The primary endpoint is the CR or VGPR rate in Cohort 1 assessed by IRC.

The following two hypotheses of zanubrutinib compared to ibrutinib in CR or VGPR rate will be tested in the analyses of the primary endpoint:

- Superiority of zanubrutinib compared to ibrutinib in the RR Analysis Set;
- Superiority of zanubrutinib compared to ibrutinib in the ITT Analysis Set.

The 2 hypotheses listed above will be tested each at a 1-sided level of 0.025 using the hierarchical fixed-sequence procedure in the following order to adjust for multiplicity: 1) superiority in the RR Analysis Set; and 2) superiority in the ITT Analysis Set. The study is positive if the superiority of zanubrutinib compared to ibrutinib in the CR or VGPR rate is demonstrated in the RR Analysis Set.

The superiority of the primary endpoint of VGPR/CR rate will be tested using the Cochran-Mantel-Haenszel (CMH) test stratified by the CXCR4 status (WHIM vs WT/missing), the prior line of therapy (1-3 vs. >3 for analyses in the Relapsed/Refractory Analysis Set; 0 vs 1-3 vs >3 for analyses in the ITT Analysis Set) and age group (<=65 vs >65) at a 1-sided significance level of 0.025. If the 1-sided p-value is less than 0.025, it will be concluded that the VGPR/CR rate in zanubrutinib is greater than the VGPR/CR rate in ibrutinib and that the primary objective is met.

The 95% confidence interval (CI) for the Mantel-Haenszel common risk difference (Mantel-Haenszel, 1959) will be constructed using a normal approximation and Sato's standard error (Sato 1989) stratified by the CXCR4 status (WHIM vs WT/missing), the prior line of therapy (1-3 vs. >3 for analyses in the Relapsed/Refractory Analysis Set; 0 vs 1-3 vs >3 for analyses in the ITT Analysis Set) and age group (<=65 vs >65).

The primary analysis of superiority in the primary endpoint will be performed in the RR Analysis Set first at least 15 months after 90% enrollment in the RR Analysis Set was completed. At the time, 90% of Cohort 1 subjects are expected to have at least 19 months of study follow up. If the superiority of the CR or VGPR rate is demonstrated in the RR analysis set, the superiority of the CR or VGPR rate will be further tested in the ITT Analysis Set.

Analyses of the best overall response (BOR) will be performed as well. BOR is defined as the best response recorded from the randomization or enrollment date to the data cutoff date or the start of new anti-cancer therapies, whichever comes first. Subjects without any post-baseline response assessment (regardless of the reason) will be considered as non-responders.

6.4.2 Secondary Efficacy Analyses

The key secondary endpoint of major response rate (MRR) by IRC will be tested only if any of the superiority tests for the primary endpoint is significant. If the primary endpoint of CR or VGPR rate is superior in the RR Analysis Set only, the key secondary endpoint of MRR will be tested for noninferiority in the RR Analysis Set at a 1-sided significance level of 0.025. If the primary endpoint of CR or VGPR rate is superior in both the RR and the ITT Analysis Sets, the key secondary endpoint of MRR will be tested for noninferiority in the RR and the ITT Analysis Set respectively at a 1-sided significance level of 0.025. The study-wide type I error will be controlled at a 1-sided 0.05 level under this testing scheme (Bretz 2009).

For the other secondary endpoints including PFS, the tests will be descriptive without multiplicity adjustment.

6.4.2.1 Key secondary endpoint: Major Response Rate

The major response rate (MRR) by IRC, defined as the proportion of subjects achieving CR, VGPR, and PR, will be tested for noninferiority of zanubrutinib compared to ibrutinib. The null and alternative hypotheses of MRR are set as follows:

H_0: MRR_A-$MRR_B \leq$ -8%

Ha: MRR_A- $MRR_B >$ -8%,

where MRR_A is the major response rate in zanubrutinib and MRR_B is the major response rate in ibrutinib.

The 95% confidence interval (CI) for the Mantel-Haenszel common risk difference (Mantel-Haenszel, 1959) will be constructed with normal approximation and standard error based on Sato (1989) with strata CXCR4 status (WHIM vs WT/missing), prior line of therapy (1-3 vs. >3 for Relapsed/Refractory analysis set analysis and 0 vs 1-3 vs >3 in ITT analysis) and age group (<=65 vs >65). If the lower bound of the CI is greater than the non-inferiority margin of -8%, the null hypothesis will be rejected, and it can be concluded that the MRR in zanubrutinib is non-inferior to the MRR in ibrutinib. In addition, as a sensitivity analysis, the Mantel-Haenszel common risk difference will also be estimated using the null variance estimator (Klingenberg, 2013).

If the lower bound of the CI is greater than 0, superiority is significant at the nominal level of 0.025 (1-sided). The superiority test of MRR is not included in the multiplicity adjustment for the study-wide type-I error (Figure 2).

As a sensitivity analysis, non-inferiority for MRR will be also tested in the per-protocol analysis set corresponding to the analysis set for which non-inferiority is significant.

MRR and the Clopper-Pearson 95% confidence interval (CI) will be reported for each arm.

Justification of the non-inferiority margin for MRR

The same non-inferiority margin will be used for both ITT and Relapsed/Refractory analysis sets.

Since no randomized trial has been conducted in WM, the ibrutinib treatment benefit over placebo can only be estimated from the single-arm ibrutinib phase 2 trial in which the MRR is 0.73 with 95% CI (0.60, 0.83). As a ≥ 50% reduction of IgM is impossible without treatment, MRR can be considered 0% in the placebo treated subject. Therefore, the lower bound of 95% CI (i.e., 60%) can be used as the treatment effect of ibrutinib over placebo (M1). A NI margin of 8% (M2) is proposed assuming 86.7% of the ibrutinib benefit over a placebo is retained. With close to 90% of the ibrutinib effect preserved, the 8% NI margin in MRR is clinically justified. The loss of the ibrutinib treatment benefit is well within the clinically acceptable range. It is worth noting that the NI of MRR will only be performed after the superiority of zanubrutinib over ibrutinib in CR/VGPR rate, which is closely correlated with MRR, has been demonstrated. Hence, a numerically higher MRR in zanubrutinib is certainly expected if the NI test of MRR is to be performed. The inclusion of the NI test of MRR is mainly based on the sample size consideration in a rare disease setting where statistical significance might not be shown in a superiority test with a clinically meaningful MRR difference (e.g., 12% higher MRR in zanubrutinib over ibrutinib) in 150 randomized relapsed/refractory subjects.

6.4.2.2 Other secondary endpoints

6.4.2.2.1 Progression-Free Survival by IRC

Progression-free survival (PFS) will be analyzed at the time of the primary analysis of VGPR/CR rate. In addition, after the primary analysis subjects will continue to be followed for PFS events until the final analysis of PFS which is scheduled at approximately 4 years after study start. During this period, the sponsor will continue to maintain trial integrity according to the Data Integrity Protection Plan (DIPP).

PFS is defined as the time (in months) from the randomization or enrollment date to the date of the earliest occurrence of disease progression (PD) or death (due to any cause), whichever occurs first. PFS will be analyzed twice: at the time of the primary analysis and approximately 4 years after the first subject was randomized.

PFS will be right-censored for subjects who meet one of the following conditions: 1) no baseline disease assessments; 2) starting a new anti-cancer therapy before PD or death; 3) PD or death immediately after more than 6 months since the last disease assessment (more than 12 months if a subject is on the response assessment schedule of every 24 weeks); and 4) alive without documentation of PD. The censoring convention generally follows the FDA Guidance for

Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics (2018) and references within the guidance document. The censoring rules are summarized in Table 1.

Table 1. Date of Progression or Censoring for Progression-free Survival

Situation	Date of Progression or Censoring	Outcome
Death or PD between the planned disease assessments	Date of death or first disease assessment showing PD, whichever occurs first	Event
Death before the first disease assessment	Date of death	Event
No baseline disease assessments	Date of randomization	Censored
New anti-cancer treatment started before or without documentation of PD or death	Date of last disease assessment prior to or on date of new anti-cancer treatment	Censored
Death or PD immediately after more than 6 months [1] from the last disease assessment	Date of the last disease assessment before death or PD	Censored
Alive without documentation of PD	Date of last disease assessment	Censored

[1] 12 months if a subject is on the assessment schedule of every 24 weeks.

Alternative censoring rules such as 1) not censoring death or PD immediately after more than 6 months from the last disease assessment and 2) ignoring new anti-cancer therapy start will be applied as sensitivity analysis at the final analysis of PFS.

The distribution of PFS, including the median and the PFS rate at selected timepoints such as 12, 18 and 24 months, will be estimated using the Kaplan-Meier method. The 95% CI for the median and the other quartiles of PFS will be estimated using Brookmeyer method (Brookmeyer and Crowley 1982). The 95% CI for the PFS rate at the selected timepoints will be estimated

using the Greenwood formula (Greenwood 1926). The duration of follow-up for PFS will be estimated using the reverse Kaplan-Meier method (Schemper and Smith 1996).

The HR (Arm A/Arm B) for PFS by IRC and its 2-sided 95% confidence interval (CI) will be estimated from a stratified Cox regression model stratified by the CXCR4 status (WHIM vs WT/missing), the prior lines of therapy (1-3 vs. >3 for the Relapsed/Refractory Analysis Set analysis and 0 vs 1-3 vs >3 in the ITT Analysis Set) and age group (<=65 vs >65) will be performed only at the final analysis of PFS. An unstratified Cox regression model will be also used to estimate the hazard ratio (HR) of the PFS in zanubrutinib compared to ibrutinib and the corresponding 95% CI at the final analysis of PFS.

The Kaplan-Meier curves for PFS will be presented for each arm. A listing of the PFS related information, e.g. the date of the progression or censoring and the corresponding reasons, will also be provided.In addition to the analysis at the time of the primary analyses of the primary endpoint, the final analysis of PFS will be performed approximately 48 months after study start. In the iNNOVATE monotherapy arm (Dimopoulos 2017), the 18-month PFS rate is estimated as 0.86. Assuming an exponential distribution for PFS and a hazard ratio of 0.8, approximately 33 events are expected to be accumulated for Cohort 1 relapsed/refractory subjects at the time of the final PFS analysis. The landmark PFS rates at 36-month for Arm A and Arm B based on the above assumptions are 0.79 and 0.74, respectively.

After the initial analysis of PFS at the time of primary analysis of the primary endpoint, the number of PFS events at the final analysis will be re-assessed based on Bagiella and Heitjan (2001). Exponential distribution will be assumed for event rate for each arm. The data observed at initial analysis will be used to determine the posterior predictive distribution. The range of projected number of PFS events will be calculated based on Bayesian prediction interval. At the final analysis of PFS, the hazard ratio will also be estimated using a Bayesian Cox regression model based on the partial likelihood and a non-informative prior. The posterior probability of hazard ratio < 1 will calculated based on the posterior distribution of log hazard ratio.

6.4.2.2.2 Duration of Response (DOR) by IRC

Duration of major response is defined as the time (in months) from the date of the earliest PR or better to the date of PD or death for any cause (whichever occurs earlier) for the subjects who achieved PR or better. The censoring convention and the analyses of duration of major response will be the same as those of PFS.

In addition, duration of VGPR or better and duration of overall response (defined as the time from date of the earliest minor response or better to PD or death due to any cause, whichever occurs earlier) will be analysed similarly as with duration of major response. Duration of responses (major response, VGPR or better, overall response) will not be compared between two arms.

6.4.2.2.3 VGPR/CR rate by Investigator

Analyses of the VGPR/CR rate will be repeated based on the investigator assessment with the same method as described in Sec 6.4.1. The concordance between the IRC and Investigator assessments in terms of VGPR/CR status and BOR will be summarized.

6.4.2.2.4 MRR by Investigator

Analyses of MRR will be repeated based on the investigator assessment with the same method as described in Sec 6.4.2.1.

6.4.2.2.5 PFS by Investigator

Analyses of PFS will be repeated based on the investigator assessment with the same method as described in Sec 6.4.2.2.1.

6.4.2.2.6 DOR by Investigator

Analyses of DOR will be repeated based on the investigator assessment with the same method as described in Sec 6.4.2.2.2.

6.4.2.2.7 Resolution of Treatment-Precipitating Symptoms

Resolution of the treatment-precipitating symptoms is defined as the absence of the symptoms that triggered the initiation of the study treatment (per the IWWM treatment guidelines, Table 2) at any point during study treatment.

The difference in the resolution of any treatment-precipitating symptoms and in the resolution of all treatment precipitating symptoms between zanubrutinib and ibrutinib will be tested using chi-square test. The number and percentage of subjects with the resolution of each and all symptoms will be summarized.

Table 2. Treatment Precipitating Symptoms

Clinical indications for initiation of therapy
Recurrent fever, night sweats, weight loss, fatigue
Hyperviscosity
Lymphadenopathy which is either symptomatic or bulky (≥5 cm in maximum diameter)
Symptomatic hepatomegaly and/or splenomegaly
Symptomatic organomegaly and/or organ or tissue infiltration
Peripheral neuropathy due to WM
Laboratory indications for initiation of therapy
Symptomatic cryoglobulinemia
Cold agglutinin anemia
Immune hemolytic anemia and/or thrombocytopenia
Nephropathy related to WM
Amyloidosis related to WM
Hemoglobin ≤10 g/dL
Platelet count $<100 \times 10^9$/L

6.4.2.2.8 Anti-Lymphoma Effect

Anti-lymphoma effect is defined as any reduction in bone marrow involvement by lymphoplasmacytoid lymphocytes and/or size of lymphadenopathy and/or splenomegaly by CT scan, at any point during study treatment.

The maximum percentage of decrease in bone marrow lymphoplasmacytoid lymphocytes will be compared using the ANOVA (analysis of variance) model with the treatment arms (arm A and B) as the group variable. The number and percentage of subjects with a decreased percentage of lymphoplasmacytoid lymphocytes after the start of the study drug will be summarized within

each treatment arm. The difference in the proportions between zanubrutinib and ibrutinib will be tested using chi-seq test.

The number and percentage of subjects with a reduction in the size of the pre-treatment lymphadenopathy and/or splenomegaly according to CT scan will be summarized within each treatment arm. The difference in the proportions between zanubrutinib and ibrutinib will be tested using chi-seq test.

6.4.3 Sensitivity Analyses

For the primary endpoint of VGPR/CR rate, an unstratified analysis will also be performed.

6.4.4 Subgroup Analyses

The primary and selected secondary endpoints will be summarized, as appropriate, in the following subgroups:

- Age group (<=65 vs. >65);
- Gender;
- Geographic region (Asia vs. Australia/New Zealand vs. Europe vs. North America);
- Prior line of therapy (0 vs. 1-3 vs. >3 and RR vs. TN);
- Baseline ECOG-PS (0 vs. ≥1);
- Baseline *CXCR4* mutation status by Sanger method (WHIM vs. WT/missing);
- Baseline IgM (≤40 g/L vs >40 g/L);
- Baseline β2 microglobulin (≤3 mg/L vs. >3 mg/L);
- Baseline Hemoglobin (≤110 g/L vs. >110 g/L);
- Baseline Platelet (≤100 10^9/L vs. >100 10^9/L);
- Baseline presence of extramedullary disease (yes vs. no);
- WM IPSS (low vs. intermediate vs. high)

When the sample size of a subgroup is too small, the subgroup may be omitted from the analyses or be combined with other subgroups. Forest plots for the subgroup analyses will be provided with unstratified rate difference and its 95% confidence interval.

6.4.5 Exploratory Efficacy Analyses

EORTC QLQ-C30

The scores at each assessment timepoint and changes from baseline in the global health status/QoL scale and the five functional scales (Physical Functioning, Role Function, Emotional Functioning, Cognitive Functioning, and Social Functioning), three Symptom scales (fatigue, pain, and nausea and vomiting), and six single items (dyspnea, insomnia, appetite loss, constipation, diarrhea, and financial difficulties) will be summarized descriptively. The percentage of subjects with clinically meaningful changes from baseline in 'global health status/QOL' and functional domains will be summarized as "improved", "stable", or "worsened".

In addition to descriptive analysis, the QLQ-C30 global health status/QOL scale scores will be compared between treatment groups in Cohort 1 using a linear mixed model for repeated measures (MMRM). The model will include the repeated measurement (including baseline) of QLQ-C30 global health status/QOL scale as dependent variable and treatment arm and randomization stratification factors, and intercept and slope of time on as fixed effects (Mallinckrodt 2018). The random subject effects will include subject random intercept and subject random slope of time. The 2 random subject effects are assumed to follow a bivariate normal distribution (with mean 0). The difference between treatment arms as well as its 95% confidence interval will be estimated, and the corresponding p-value will be presented.

EQ-5D-5L

The EQ-5D-5L comprises a descriptive system and an EQ Visual Analogue scale (EQ VAS) with following 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The EQ VAS records the respondent's self-rated health on a 0 to 100 scale, with 100 labelled 'the best health you can imagine' and 0 'the worst health you can imagine'.

EQ VAS will be summarized descriptively. Descriptive statistics will also be performed including the number and percentage of subjects reporting each level of problem on each dimension of the EQ-5D

Medical resource utilization, including the number of hospitalizations, the length of hospital stay, and supportive care (e.g., transfusions, growth factor support, intravenous antibiotics), will be summarized at each cycle (including Screening and C1D1) for each study arm in Cohort 1.

Correlation between the trough plasma zanubrutinib concentration and the efficacy endpoints may be explored in the PK analysis set.

[redacted].

Additional analyses may be performed by combining the data from Arms A and C to assess the efficacy of zanubrutinib, regardless of the *MYD88* mutation status.

6.5 SAFETY ANALYSES

All safety analyses will be performed in the Safety Analysis Set.

Safety will be assessed by the monitoring and recording of all adverse events (AEs) graded by NCI-CTCAE v4.03. Laboratory values (complete blood count [CBC], serum chemistry, coagulation, and urinalysis), vital signs, physical exams, and ECGs findings will also be used in determining the safety. Descriptive statistics will be used to summarize the safety data by treatment arm. For all safety analyses, subjects will be analyzed according to the treatment arm corresponding to the actual treatment received. The safety data will be summarized by study drug (combining Arm A and Arm C for zanubrutinib) as well.

6.5.1 Extent of Exposure

Extent of exposure to study drug will be summarized descriptively as the number of cycles received (number and percentage of subjects), duration of exposure (days), cumulative total dose received per subject (mg), actual dose intensity (mg/day) and relative dose intensity (%).

The number and percentage of subjects with dose reductions, dose interruption, and drug discontinuation due to AEs will be summarized. The cycle in which the first dose reduction/interruption occurred will be summarized using descriptive statistics. The frequency of dose reductions and dose interruptions will be summarized by categories.

The actual dose intensity is defined as the actual cumulative dose (mg) taken based on the total dose per day divided by the duration of treatment (days).The relative dose intensity (RDI) is defined as the ratio of the actual dose intensity to the planned dose intensity in percentage.

Subject data listings will be provided for all dosing records and for the calculated summary statistics.

In addition, time to treatment failure will calculated as time from randomization to discontinuation of study treatment due to any reason. Time to treatment failure will be censored at the data cutoff for the subjects who did not discontinue study treatment. The hazard ratio for time to treatment failure and its 95% CI will be estimated from a stratified Cox regression using the randomization stratification factors and age group. The Kaplan-Meier method will be used to estimate the distribution of time to treatment failure for each treatment group.

6.5.2 Adverse Events

The AE verbatim descriptions (investigator terms from the CRF) will be classified into standardized medical terminology using Medical Dictionary for Regulatory Activities (MedDRA®). Adverse events will be coded to MedDRA® (Version 18.1 or higher) lower level term closest to the verbatim term. The linked MedDRA® preferred term (PT) and primary system organ class (SOC) are also captured in the database.

A treatment-emergent adverse event (TEAE) is defined as an AE with an onset time or increase in severity level on or after the first dose of study drug and within 30 days after the last dose of study drug or prior to the initiation of a new anti-cancer therapy, whichever occurs first. A treatment-related adverse event is an AE that is assessed by the investigator as related to the study drug or with missing assessment of the causal relationship.

TEAEs will be summarized based on the number and percentage of subjects experiencing AEs by MedDRA SOC and PT. A subject reporting the same AE more than once will be counted only once when calculating the incidence 1) within a given SOC, and 2) within a given SOC and PT combination. The maximum CTCAE toxicity grade and strongest causal relationship to the study drug for such events will be used.

An overall summary of TEAEs will include the number and percentage of subjects with at least one:

- TEAE;
- treatment-related TEAE;
- grade 3 or higher TEAE;
- treatment-related grade 3 or higher TEAE;
- fatal TEAE;

- treatment-related fatal TEAE;
- serious TEAE;
- treatment-related serious TEAE;
- TEAE leading to treatment discontinuation;
- treatment-related TEAE leading to treatment discontinuation;
- TEAE leading to dose reduced;
- treatment-related TEAE leading to dose reduced;
- TEAE leading to dose interruption;
- treatment-related TEAE leading to dose interruption.

The summaries of the TEAEs will be provided by PT only, by SOC and PT, and by SOC, PT and maximum severity.

The adverse events of special interest (AESIs) will be defined and summarized by AESI category and PT for the following:

- all AESIs;
- grade 3 or higher AESIs;
- serious AESIs;
- AESIs by maximum severity.

Time to first AESI along with cumulative event rates at milestone timepoints as well as subject incidence in 3-month intervals over time may be presented for selective AESIs.

The following AEs of hepatic events will be summarized by PT:

- All AEs of hepatic events;
- AEs of hepatic events grade 3 or higher;
- Serious AEs of hepatic events;
- AEs of hepatic events by maximum severity.

A summary of the number of deaths and the cause of death, classified by deaths within 30 days of last dose of study drug and deaths more than 30 days after the last dose, will be provided.

Listings of deaths, AEs, serious AEs, AEs leading to death, and AEs leading to dose modification or discontinuation of the study drug will be provided.

6.5.3 Laboratory Values

The actual value and change from baseline for selected hematology, serum chemistry, coagulation, and urinanalysis parameters will be summarized at each scheduled visit. A list of parameters of interest is provided in Table 3. Laboratory parameters that are graded in NCI-CTCAE v4.03 will be summarized by the CTCAE grade.

A summary of the number and percentage of subjects with grade 3 or higher toxicity will be provided for selected laboratory parameter of interest. In the summary of laboratory parameters by CTCAE grade, parameters with CTCAE grading in both high and low directions (e.g., calcium, glucose, magnesium, potassium, sodium) will be summarized separately. Shift tables assessing the toxicity grade at baseline versus the worst post-baseline toxicity grade recorded will be presented.

Table 3. Laboratory Parameters of Interest

Hematology	Serum Chemistry		Coagulation
Hemoglobin (decrease)	Alanine transaminase (ALT) (increase)	Albumin (decrease)	Activated partial thromboplastin time (aPTT) (increase)
Platelets (decrease)	Aspartate transaminase (AST) (increase)	Uric Acid (increase)	International Normalized Ratio (INR) (increase)
WBC (increase, decrease)	Alkaline Phosphatase (increase)	Sodium (increase, decrease)	
Absolute Neutrophil Count (ANC, decrease)	Total Bilirubin (increase)	Phosphorus (decrease)	

Absolute Lymphocyte Count (increase, decrease)	Creatinine (increase)	Potassium (increase, decrease)
	Calcium (increase, decrease)	Magnesium (increase, decrease)
	Glucose (increase, decrease)	

For hypocalcemia and hypercalcemia, serum calcium will be corrected using the following formula:

> Corrected calcium = Serum calcium + 0.8 * (4 – serum albumin) where serum calcium is recorded in mg/dL and serum albumin is recorded in g/dL.

Box-whisker plots will be generated for parameters of interest.

Hematology, serum chemistry, coagulation, and urinalysis results for each subject will be presented in data listings. A listing of all grade 3 or higher laboratory values will also be provided.

Incidence of subjects who met one or more of the Hy's law criteria will be summarized. A listing of subjects that met one or more of the Hy's law criteria will be generated.

6.5.4 Vital Signs

Descriptive statistics for vital sign parameters (systolic and diastolic blood pressure, heart rate, temperature, weight) and changes from baseline will be presented by visit. A listing by subject and visit will be provided as well.

6.5.5 Electrocardiograms (ECG)

Actual value and change from baseline for the PR, QT, and QTc-Fridericia intervals will be summarized. If triplicate readings are recorded (e.g. Screening), the average of the readings for the visit will be used for the summary.

The number and percentage of subjects satisfying the following QT and QTcF conditions at any time post-baseline will be summarized:

- > 450, > 480, or > 500 msec;
- ≤ 30 msec increase from baseline, > 30 and ≤ 60 msec increase from baseline, or > 60 msec increase from baseline.

A listing of abnormal ECG by subject and visit will be provided.

6.5.6 ECOG

ECOG performance status will be summarized at each visit. Shift tables assessing the ECOG performance status at baseline versus the worst post-baseline performance status will be presented.

6.6 PHARMACOKINETIC ANALYSES

Population PK analysis will be carried out to include plasma concentrations from this trial in an existing model. PK parameters such as C_{min} will be summarized, and additional PK parameters such as apparent clearance of the drug from plasma (CL/F) and area under the plasma concentration-time curve ($AUC_{[0-24]}$) may be derived from the population PK analysis if supported by data.

Exposure-response (efficacy or safety endpoints) analysis may be carried out if supported by data.

The impact of plasmapheresis on zanubrutinib exposure will be estimated by comparing the AUC of zanubrutinib in subjects with or without the plasmapheresis procedure.

Further PK analyses will be described in a separate analysis plan.

7 INTERIM ANALYSIS

An interim analysis, which occurs approximately 6 months after the first 50 relapsed/refractory subjects are randomized from Cohort 1, will be implemented in the study. This analysis is for futility only. Early stopping for efficacy will not be considered for this interim analysis.

Therefore, only nominal alpha spending 0.00001 will be performed due to the interim analysis.

No recruitment stop is planned for this interim analysis. If futility is demonstrated in the interim analysis, enrollment will be halted and safety and efficacy data will be further evaluated before making the decision of stopping the study permanently. Otherwise, the enrollment will be continued to 150 relapsed/ refractory subjects in Cohort 1.

Details of the interim analysis is provided in an interim analysis SAP.

8 CHANGES IN THE PLANNED ANALYSIS

8.1 TIMING OF THE PRIMARY EFFICACY ANALYSIS

In the protocol amendment 4.0, it is specified that the primary efficacy analysis will be conducted approximately 12 months after the last relapsed/refractory subject is randomized. In this version of SAP, the primary efficacy analysis is planned to occur at least 15 months after 90% enrollment in the RR Analysis Set was completed. At the time, 90% of subjects in Cohort 1 are expected to have at least 19 months of median study follow-up. This is considered sufficient to demonstrate an improvement of clinical benefit of zanubrutinib over ibrutinib.

8.2 TESTING STRATEGY FOR THE PRIMARY ENDPOINT FAMILY AND THE KEY SECONDARY ENDPOINT FAMILY

Testing strategy and multiplicity adjustment for the primary and key secondary endpoint hypotheses are added to Section 6.2.4 to provide a formal and rigorous testing framework based on Bretz (2009). As a result, the family-wise error rate (FWER) within the primary hypotheses family is controlled at a 0.025 level (1-sided) and the FWER within both primary and secondary hypotheses (i.e. the study-wise error rate) is controlled at a 0.05 level (1-sided).

8.3 AGE GROUP AS COVARIATE IN STRATIFIED ANALYSIS OF THE PRIMARY AND SECONDARY ENDPOINTS

Age group (<=65 vs >65) is added as additional covariate in the stratified analysis of the primary and secondary endpoints because it is considered as an important prognostic factor and one of

five variables included in the final predictive model in IPSS-WM (Morel, 2009). The cutoff of age group is based on IPSS age variable cutoff for low risk and intermediate/high risk subjects.

8.4 TIMING OF THE FINAL ANALYSIS OF PFS

In this version of SAP, it is added that the final analysis of PFS will be performed approximately 48 months after study start as described in Sec 6.4.2.2. Assuming an exponential distribution for PFS, the control arm median PFS of 84 months, a hazard ratio of 0.8, approximately 33 events are expected to be accumulated for Cohort 1 relapsed and refractory subjects at the time of the final PFS analysis. The estimated landmark PFS rates at 36-month for Arm A and Arm B are 0.79 and 0.74, respectively based on the above assumptions.

9 REFERENCES

Bretz, F., Maurer, W., and Paul Gallo, P. (2009) Discussion of "Some Controversial Multiple Testing Problems in Regulatory Applications" by H. M. J. Hung and S.-J. Wang, Journal of Biopharmaceutical Statistics, 19:1, 25-34.

Brookmeyer, R. and Crowley, J. (1982) *A confidence interval for the median survival time*. Biometrics, 38, 29-41.

Dimopoulos MA, et al. Treatment recommendations for patients with Wanldenstrom macroglobulinemia (WM) and related disorders: IWWM-7 consensus. Blood. 2014. 124(9):1404-1411.

Dimopoulos MA, et al. Ibrutinib for patients with rituximab-refractory Waldenström's macroglobulinaemia (iNNOVATE): an open-label substudy of an international, multicentre, phase 3 trial. Lancet Oncology. 2017.18 (2):241-250.

EMA guidance on points to consider on switching between superiority and noninferiority, 2000

Fayers, P. M., N. K. Aaronson, K. Bjordal, M. Groenvold, D. Curran, A. Bottomley, and on behalf of the EORTC Quality of Life Group. 2001. The EORTC QLQ-C30 Scoring Manual. 3rd ed. Brussels: European Organization for Research and Treatment of Cancer.

Food and Drug Administration Center for Drug Evaluation and Research (CDER). FDA Guidance for Industry: Non-Inferiority Clinical Trials to Establish Effectiveness. (2016).

Food and Drug Administration Center for Drug Evaluation and Research (CDER) & Center for Biologics Evaluation and Research (CBER). FDA Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. (2018)

Greenwood, M. (1926) The natural duration of cancer. Public health and medical subjects, 33: 1-26.

Herdman, M et al, Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res (2011) 20:1727–1736.

Mallinckrod†, C. H., Lane, P. W., Schnell, D., Peng, Y., & Mancuso, J. P. (2008). Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. *Drug Information Journal*, *42*(4), 303–319.

Mantel, N. and Haenszel, W. (1959), Statistical Aspects of Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute. 22, 719–748.

Morel, P., Duhamel, A., Gobbi, P., Dimopoulos, M. A., Dhodapkar, M. V., McCoy, J., Crowley, J., Ocio, E. M., Garcia-Sanz, R., Treon, S. P., Leblond, V., Kyle, R. A., Barlogie, B., & Merlini, G. (2009). International prognostic scoring system for Waldenström macroglobulinemia. Blood, 113(18), 4163-4170.

NCI-CTCAE v4.03
http://www.oncology.tv/SymptomManagement/NationalCancerInstituteUpdatesCTCAEtov403.aspx. Accessed 23 May 2019.

Owen et al. Response assessment in Waldenström macroglobulinaemia: update from the VIth International Workshop. Br J Haematol. 2013 Jan;160(2):171-6.

Owen et al 2013 and NCCN Guidelines, Lymphoplasmacytic Lymphoma/Waldenström's Macroglobulinemia. IWWM. 2015: v2

Sato, T. (1989). On the Variance Estimator of the Mantel-Haenszel Risk Difference. Biometrics. 45, 1323–1324, letter to the editor.

Schemper, M, and Smith, T.L. (1996). A note on quantifying follow-up in studies of failure time. Controlled Clinical Trials, 17:343-346

Szende A, Oppe M, Devlin N. (Eds.) EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Springer 2007.

The EuroQol Group, 1990. EuroQol – a new facility for the measurement of health-related quality of life. Health Policy, 16, pp.199-208.

Treon SP et al. Ibrutinib in previously treated Waldenstrom's Macroglobulinemia. N Engl J Med. 2015 Apr 9;372(15):1430-40.

10 APPENDIX

10.1 APPENDIX A: IMPUTATION OF MISSING OR PARTIALLY MISSING DATES

In general, missing or partial dates will not be imputed as data level. The following rules will apply for the specific analysis and summary purposes mentioned below only.

A.1 Prior/Concomitant Medications/Procedures

When the start date or end date of a medication is partially missing, the date will be imputed to determine whether the medication is prior or concomitant. The following rules will be applied to impute partial dates for medications:

If start date of a medication is partially missing, impute as follows:

- If both month and day are missing, then set to January 01
- If only day is missing, then set to the first of the month

If end date of a medication is partially missing, impute as follows:

- If both month and day are missing, then set to December 31
- If only day is missing, then set to last day of the month

If start date or end date of a medication is completely missing, do not impute.

A.2 Adverse Events

The imputation rule for the safety analyses will be used to address the issues with partial dates. When the start date or end date of an AE is partially missing, the date will be imputed to determine whether the AE is treatment-emergent. When in doubt, the AE will be considered treatment emergent by default. The following rules will be applied to impute partial dates for AEs:

If start date of an AE is partially missing, impute as follows:

- If both month and day are missing, then the imputed day and month will be January 01 or the first dosing date if they have the same year, whichever is later.
- If only day is missing, then the imputed day will be the first day of the month or the first dosing date if they have the same month and year, whichever is later
- If start date is completely missing, the imputed day will be the first dosing date as long as AE end date is not before the first dosing date.

If end date of an AE is partially missing, impute as follows:

- If both month and day are missing, then set to December 31
- If only day is missing, then set to last day of the month

If end date is completely missing, do not impute.

A.3 Deaths

In case complete death dates are not recorded, impute as follows:

- If both month and day are missing, then the imputed month and day will be 01Jan or the last date of the last date of subject known to be alive + 1, whichever is later.
- If only day is missing, the death will be assumed to be on the first day of the month or the last date of subject known to be alive +1, whichever is later.

A.4 New Anti-cancer therapy

If the start day of a subsequent anti-cancer therapy is incomplete or missing, impute as follows:

- If both month and day are missing, then the imputed month and day will be 01Jan or the last day of the month for the last adequate disease assessment if they have the same year.
- If only day is missing, then the imputed day will be the first day of the month.

A.5 Diagnosis

If a diagnosis date is partially missing, impute as follows:

- If both month and day are missing, then set to January 01
- If only day is missing, then set to the first of the month

If a diagnosis date is completely missing, do not impute.

A.6 Prior Therapy/Response to Prior Therapy

If a prior therapy or response to prior therapy date is partially missing, impute as follows:

If only day is missing, then set to the 15th of the month

10.2 APPENDIX B: WM INTERNATIONAL PROGNOSTIC SCORING SYSTEM (IPSS)

Score:

- Age > 65 yr
- Hemoglobin ≤ 11.5 g/dL
- Platelet count ≤ 100 x 10^9/L
- β2-microglobulin > 3 mg/L
- Monoclonal IgM concentration (M-protein by SPEP)> 7.0 g/dL

IPSS:

- Low 0 or 1 (except age) score
- Intermediate age or 2 scores
- High ≥ 3 scores

Source: Morel (2009)