STATISTICAL ANALYSIS PLAN

Study Protocol Number: **BGB-3111-305**

Study Protocol Title: **A Phase 3, Randomized Study of Zanubrutinib (BGB-3111) Compared with Ibrutinib in Patients with Relapsed/Refractory Chronic Lymphocytic Leukemia or Small Lymphocytic Lymphoma**

Date: March 12, 2021

Version: Final 1.0

NCT Number NCT03734016

SIGNATURE PAGE

[redacted]

Director, Biostatistics

Signature [redacted]

Date: __

Approval

[redacted]

Vice President, Global Statistics and Data Science

Signatur [redacted]

Date: __

[redacted]

Chief Medical Officer

Signature [redacted]

Date: ___

[redacted]

Medical Director

Signature [redacted]

Date: ___

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

Abbreviation	**Term**
ADaM	Analysis data model
AE	Adverse event
ALT	Alanine aminotransferase
aPTT	Activated partial thromboplastin time
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’s tyrosine kinase
CLL	Chronic lymphocytic leukemia
C_{max}	Maximum observed plasma concentration
CR	Complete response
CRi	Complete response with incomplete bone marrow recovery
CSR	Clinical study report

CT	Computed tomography
DBP	Diastolic blood pressure
DIPP	Data integrity protection plan
DOR	Duration of response
eCRF	Electronic case report form
ECG	Electrocardiogram
ECOG	Eastern cooperative oncology group
EDC	Electronic data capture
LDH	Lactate dehydrogenase
LLN	Lower limit of normal
MAR	Missing at random
MedDRA	Medical Dictionary for Regulatory Activities
MRD	Minimal residual disease
NCI CTCAE	National Cancer Institute Common Toxicity Criteria for Adverse Events
ORR	Overall response rate
OS	Overall survival
PD	Progressive disease
PFS	Progression-free survival

PK	Pharmacokinetic
PR	Partial response
PR-L	Partial response with lymphocytosis
PT	Preferred term
Q1, Q3	First quartile, third quartile
QD	Once daily
QT	Electrocardiographic interval
SAE	Serious adverse event
SAP	Statistical analysis plan
SBP	Systolic blood pressure
SE	Standard error
SLL	Small lymphocytic lymphoma
SOC	System organ class
SPD	Sum of products of the perpendicular diameters
SD	Stable disease
$t_{1/2}$	Terminal half-life
TEAE	Treatment-emergent adverse event
t_{max}	Time to maximum observed plasma concentration

TTR	Time to response
ULN	Upper limit of normal
WHO-DD	World Health Organization Drug Dictionary

1 INTRODUCTION

This statistical analysis plan (SAP) describes the detailed plans for the analysis of safety and efficacy data for study BGB-3111-305: A Phase 3, Randomized Study of Zanubrutinib (BGB-3111) Compared with Ibrutinib in Patients with Relapsed/Refractory Chronic Lymphocytic Leukemia or Small Lymphocytic Lymphoma. This SAP describes analyses for Europe and other regions outside of the United States and China; analyses for the United States and China will be covered in a separate SAP.

This document is based on the protocol version 3.0 dated 31JAN2020. The analysis plan for the pharmacokinetic and exploratory biomarker analyses are not included in this SAP.

2 STUDY OVERVIEW

This is a Phase 3, randomized open-label study of zanubrutinib versus ibrutinib in approximately 600 patients with relapsed/refractory chronic lymphocytic leukemia (CLL) or small lymphocytic lymphoma (SLL). The primary efficacy endpoint is overall response rate (ORR), defined as a best overall response of partial response (PR) or higher, per investigator assessment. Disease response will be assessed per the “modified” 2008 International Workshop on Chronic Lymphocytic Leukemia guidelines (Hallek et al 2008), with modification for treatment-related lymphocytosis (Cheson et al 2012) for patients with CLL, and per Lugano Classification for non-Hodgkin lymphoma (Cheson et al 2014) for patients with SLL. Rate of response for partial response with lymphocytosis (PR-L) or higher will be assessed as a secondary efficacy endpoint.

The study is broken into three periods for every patient:

- Screening Period
- Treatment Period
- Post-Treatment Period
    - End of Treatment Visit
    - Long-Term Follow-up
    - Survival Follow-up

Patients will be randomized in a 1:1 manner to one of the following treatment arms:

- Arm A: Zanubrutinib 160 mg orally twice daily
- Arm B: Ibrutinib 420 mg orally once daily

Randomization will be stratified by age (<65 years versus ≥ 65 years), geographic region (China versus non-China), refractory status (yes or no), and del17p/TP53 mutation status (present or absent).

Study treatment will be open-label and should continue until disease progression is confirmed by independent central review. The study duration is estimated to be approximately 51 months.

Assessments to be performed during the study include pharmacokinetics of zanubrutinib; disease-related constitutional symptoms; physical examination of the liver, spleen and lymph nodes; computed tomography (CT) or protocol allowed alternative imaging modality scan of neck, chest, abdomen and pelvis with contrast; bone marrow examination at screening, for patients with baseline marrow involvement who meet clinical and laboratory criteria for CR or CRi and at time of suspected cytopenic progression; patient-reported outcomes (PRO; European quality of life 5-dimensions 5-levels health questionnaire [EQ-5D-5L], European Organisation for Research and Treatment of Cancer quality of life cancer core questionnaire [EORTC QLQ-C30], and, in select countries, an optional device-based evaluation of quality of life and activity that patients may consent to); laboratory studies; bone marrow examination; genetic alterations in the tumor cells (eg, del 17p, TP53, del 11q, del 13q, trisomy 12 and immunoglobulin variable region heavy chain [IGHV] mutation analysis).

Assessments of safety will include adverse events (AEs), serious adverse events (SAEs), clinical laboratory tests, physical examinations, electrocardiograms, ECOG performance status, and vital signs. Adverse events will be graded for severity per the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) version 4.03 and the Grading Scale for Hematologic Toxicities in CLL Studies according to Hallek 2008. An independent Data Monitoring Committee will periodically monitor safety data and perform the interim efficacy analysis.

3 STUDY OBJECTIVES

3.1 Primary Objective

- To compare the efficacy of zanubrutinib versus ibrutinib as measured by ORR determined by investigator

3.2 Secondary Objectives

- To compare the efficacy of zanubrutinib versus ibrutinib as measured by:
  - Progression-free survival (PFS) determined by investigator assessment and independent central review
  - ORR determined by independent central review
  - Duration of response (DOR) determined by independent central review
  - DOR determined by investigator assessment
  - Time to treatment failure
  - Rate of PR-L or higher determined by independent central review
  - Overall survival
  - Health Related Quality of Life (HRQoL)
- To compare the safety of zanubrutinib versus ibrutinib

3.3 Exploratory Objectives

- To evaluate the correlation between clinical outcomes (eg, ORR, PFS, DOR, overall survival, rate of partial response) and the prognostic and predictive biomarkers, including minimal residual disease (MRD)
- To characterize the pharmacokinetics (PK) of zanubrutinib

4 STUDY ENDPOINTS

4.1 Primary Endpoint

The primary endpoint is ORR (PR or higher, defined as CR/CRi + PR + nodular PR) per investigator assessment using the "modified" 2008 IWCLL guidelines (Hallek et al 2008) with modification for treatment-related lymphocytosis (Cheson et al 2012) for patients with CLL and per Lugano Classification for non-Hodgkin lymphoma (NHL) (Cheson et al 2014) for patients with SLL.

4.2 Secondary Endpoints

Key Secondary Endpoints:

The key secondary endpoints are PFS per investigator assessment and atrial fibrillation/flutter incidence.

PFS is defined as the time from randomization to the date of first documentation of disease progression or death, whichever comes first.

For atrial fibrillation/flutter incidence, patients will be considered as having an atrial fibrillation/flutter event if they have a treatment-emergent AE of either "atrial fibrillation" or "atrial flutter".

Other Secondary Endpoints:

- ORR by independent central review
- PFS by independent central review
- DOR, defined as the time from the date that response criteria are first met to the date that disease progression is objectively documented or death, whichever comes first, determined by independent central review
- DOR by investigator assessment
- Time to treatment failure, defined as the time from randomization to discontinuation of study drug due to any reason
- Rate of PR-L or higher, defined as the proportion of patients who achieve a CR/CRi + PR + nodular PR + PR-L determined by independent central review

- Rate of PR-L or higher determined by investigator assessment
- Overall survival, defined as the time from randomization to the date of death due to any cause
- HRQoL measured by the EQ-5D-5L and EORTC QLQ-C30 questionnaires
- Safety parameters, including AEs, SAEs, clinical laboratory tests, physical exams, and vital signs

4.3 Exploratory Endpoints

- Correlation between clinical outcomes (eg, ORR, PFS, DOR, OS) and the prognostic and predictive biomarkers
- MRD negativity rate
- PK parameters
- Self-administered Activity and Quality of Life questionnaire

5 SAMPLE SIZE CONSIDERATIONS

The sample size calculation is based on the primary efficacy analyses for the primary endpoint of ORR per investigator assessment. Assuming a response ratio (zanubrutinib arm / ibrutinib arm) of 1.03 (72% / 70%), 600 patients will provide more than 90% power to demonstrate the noninferiority of zanubrutinib to ibrutinib at the noninferiority margin of 0.8558 (response ratio) and a 1-sided alpha level of 0.025 when there is 1 interim analysis at 69% information fraction. The response rate for ibrutinib is approximated from published clinical data (Byrd et al 2019).

Assuming a hazard ratio (HR) of 0.9 (zanubrutinib arm / ibrutinib arm), 205 PFS events are required to achieve 80% power at a 1-sided alpha of 0.025 to demonstrate the noninferiority of zanubrutinib to ibrutinib at the noninferiority margin of 1.3319 (HR) for the key secondary endpoint of PFS per investigator assessment. If the 600 patients are randomized in a 1:1 ratio to the 2 arms over a 24-month period, including a 9-month ramp-up period before reaching peak enrollment of 33 patients/month with a 0.0017/month hazard rate for dropout, 205 PFS events are expected to accumulate in 45 months from study start. A median PFS of 47 months for ibrutinib and an exponential distribution for PFS are also assumed.

6 STATISTICAL METHODS

6.1 ANALYSIS SETS

The Intent to Treat (ITT) Analysis Set includes all randomized patients.

The Safety Analysis Set includes all patients who received any dose of study drug. The Safety Analysis Set will be used for all safety analyses.

The Per-protocol Analysis Set includes patients who received any dose of study drug and had no critical protocol deviation. Categories of critical protocol deviations are defined in Section 6.3.2.

Criteria for exclusion from the Per-protocol Analysis Set will be determined and documented before the database lock for the interim analysis of the primary endpoint.

The PK Analysis Set includes all zanubrutinib-treated patients who have at least one post dose drug concentration measured.

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.

If an event date (eg, AE onset date) is partial or missing, the date will appear partial or missing in the listings.

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 drug (study day 1). For patients not dosed, the randomization date will be used instead of the first dose date. For assessments conducted on or after the date of first dose of study drug, study day will be calculated as (assessment date – date of first dose of study drug + 1). For assessments conducted before the date of the first dose of study drug, study day is calculated as (assessment date – date of first dose of study drug). There is no study day 0.

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

Baseline: the baseline value or assessment is defined as the last non-missing value or assessment before the first dose of study drug or before randomization for patients not dosed.

Postbaseline: a postbaseline value or assessment is defined as a value or assessment after the first dose of study drug or after randomization for patients not dosed.

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. Missing data for the health-related quality-of-life (HRQoL) data will be handled according to each PRO instrument manual (Fayers & Machin, 2000 in The EORTC QLQ-C30 (Third Edition), 2001; https://euroqol.org/publications/user-guides/).

Missing longitudinal data analyses may employ the missing-at-random (MAR) assumption such as described for QLQ-C30 scores over time in Section 6.4.2.3.9.

Time-to-event data analyses will have associated censoring rules as described for PFS in Section 6.4.2.1.

When summarizing categorical variables, patients 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 patients with missing data is presented.

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

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.

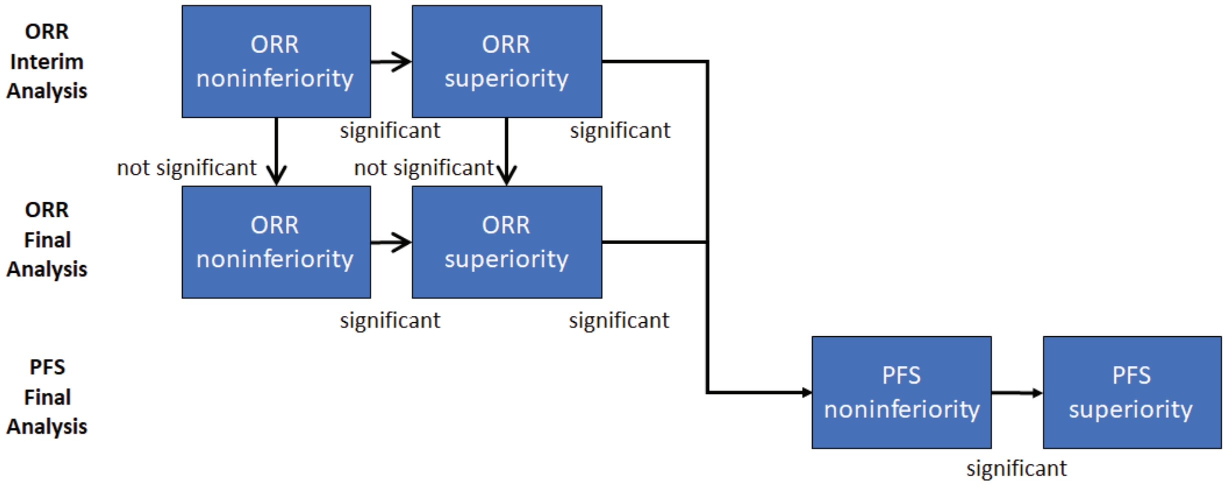
6.2.4 Multiplicity Adjustment

To control the study-wide type I error, individual significance levels will be adjusted for the tests of the primary endpoint of ORR per investigator assessment (noninferiority and superiority), and the key secondary endpoint of PFS per investigator assessment (noninferiority and superiority). Multiplicity due to multiple endpoints and multiple tests will be handled per the graphical approach by Maurer and Bretz (2013) utilizing fixed sequence hierarchical testing. Under this procedure, secondary endpoints will be tested only if the primary endpoint is significant.

If the noninferiority of ORR per investigator assessment is statistically significant, the key secondary endpoint of atrial fibrillation/flutter incidence will be tested at the interim and final analyses of ORR with the same 1-sided significance levels as ORR but will be tested separately from the fixed sequence hierarchical testing.

Hypothesis testing will be performed according to the multiplicity adjustment flowchart shown in Figure 1.

Figure 1. Flowchart for the Multiplicity Adjustment



Per the testing sequence in Figure 1, the study-wide 1-sided significance level of 0.025 will be passed to subsequent hypothesis tests in the sequence and will be distributed for each of the following potential analysis timepoints based on the known correlation of the interim and final test statistics and corresponding alpha spending function:

- interim analysis of ORR
  - noninferiority of ORR per investigator assessment is tested at a 1-sided significance level of 0.005 based on the O'Brien-Fleming alpha spending function

with an information fraction of 64% (415 divided by 652, total number of randomized patients); if this is not statistically significant, do not conduct any of the remaining hypothesis tests at this analysis timepoint and continue to the final analysis of ORR

  - if the noninferiority of ORR per investigator assessment is statistically significant, the superiority of ORR per investigator assessment will be tested at a 1-sided significance level of 0.005 based on the O'Brien-Fleming alpha spending function with an information fraction of 64%; if this is not statistically significant, continue to the final analysis of ORR for additional hypothesis testing starting with the superiority of ORR per investigator assessment
  - if the superiority of ORR per investigator assessment is statistically significant at the interim analysis, PFS per investigator assessment will be compared between the two treatment arms for descriptive purposes only, but not for statistical inference (for either non-inferiority or superiority) at this interim analysis. A 1-sided 0.00001 significance level will be spent to account for the increased false positive rate due to this descriptive analysis.

- final analysis of ORR
  - if the noninferiority of ORR per investigator assessment was not statistically significant at the interim analysis of ORR,
    1. noninferiority of ORR per investigator assessment will be tested at a 1-sided significance level of 0.0235; if this is not statistically significant, the study does not meet the primary objective and no additional hypothesis testing will be performed in this study
    2. if the noninferiority of ORR per investigator assessment is statistically significant, the superiority of ORR per investigator assessment will be tested at a 1-sided significance level of 0.0235; if this is not statistically significant, no additional hypothesis testing will be performed in this study
    3. if the superiority of ORR per investigator assessment is statistically significant, PFS per investigator assessment will be compared between the two treatment arms for descriptive purposes only, but not for statistical

        inference (for either non-inferiority or superiority). A 1-sided 0.00001 significance level will be spent to account for the increased false positive rate due to this descriptive analysis.

  - if the noninferiority of ORR per investigator assessment was statistically significant but the superiority of ORR per investigator assessment was not significant at the interim analysis of ORR
    1. the superiority of ORR per investigator assessment will be tested at a 1-sided significance level of 0.0235; if this is not statistically significant, no additional hypothesis testing will be performed in this study
    2. if the superiority of ORR per investigator assessment is statistically significant, PFS per investigator assessment will be compared between the two treatment arms for descriptive purposes only, but not for statistical inference (for either non-inferiority or superiority). A 1-sided 0.00001 significance level will be spent to account for the increased false positive rate due to this descriptive analysis.
- final analysis of PFS
  - if the superiority of ORR per investigator assessment is statistically significant at either the interim or final analysis of ORR, PFS per investigator assessment will be followed further until 205 PFS events per investigator assessment have occurred for the final PFS analysis; noninferiority of PFS per investigator assessment will be first tested at a 1-sided significance level of 0.02498
  - if the noninferiority of PFS per investigator assessment is statistically significant, the superiority of PFS per investigator assessment will be tested at a 1-sided significance level of 0.02498.

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 patients' 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.

Though the study is open-label, access to study data is controlled for specific sponsor study team members overseeing the conduct of the study or analyzing study data. The sponsor will not have access to aggregated efficacy or safety data summaries by treatment arm according to the study's Data Integrity Protection Plan (DIPP).

6.3 Patient Characteristics

6.3.1 Patient Disposition

The following patient disposition information will be summarized by treatment arm:

- Number of randomized patients
- Number (%) of treated patients
- Number (%) of treated patients still receiving treatment as of the data cutoff
- Number (%) of treated patients who discontinued treatment and reasons for treatment discontinuation
- Number (%) of randomized patients still on study as of the data cutoff
- Number (%) of randomized patients who discontinued study and reasons for study discontinuation

Study follow-up time is defined as the time from the randomization to the death date or the end of study date for the patients who discontinued from study (whichever occurs first), or the data cutoff date for patients still on study. Study follow-up time will be summarized descriptively.

6.3.2 Protocol Deviations

Important protocol deviations are protocol deviations designated as "Major" or "Significant" per the China and the Rest of the World protocol deviation data, respectively.

Critical protocol deviations will also be identified and used to define the Per-protocol Analysis Set. Critical protocol deviations will be identified before the database lock. The critical protocol deviations will include but are not limited to selected protocol deviations from the following categories:

- Patient randomized even though he/she did not satisfy study eligibility criteria
- Patient developed study drug withdrawal criteria but was not withdrawn
- Patient received wrong study treatment or incorrect dose
- Patient received prohibited concomitant treatment

Important protocol deviations and critical protocol deviations will be summarized by deviation category and by treatment arm. COVID-19-related protocol deviations will be summarized. A listing will also be provided.

6.3.3 Randomization Stratification Factors

The number of patients with each of the IRT randomization stratification factors will be summarized by treatment arm. The randomization stratification factors include age (< 65 years versus ≥ 65 years), geographic region (China versus non-China), refractory status (yes or no), and del17p/TP53 mutation status (present or absent). In addition, stratification errors (defined as IRT-based strata vs clinical data-based strata inconsistencies) will be summarized.

6.3.4 Demographics and Other Baseline Characteristics

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

- Age (years);
- Age group (< 65 vs ≥ 65 years);
- Sex;
- Race and ethnicity;
- Geographic region (Asia vs Australia/New Zealand vs Europe vs North America);
- Height (cm), weight (kg), and body mass index (BMI, kg/m^2);
- Eastern Cooperative Oncology Group (ECOG) performance status.

6.3.5 Disease History and Baseline Disease Characteristics

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

- Time since initial CLL/SLL diagnosis;
- Disease type (CLL vs SLL);
- Disease stage (Binet stage for CLL; Ann Arbor stage for SLL) at study entry;
- Bulky disease (any target lesion longest diameter ≥ 5 cm; ≥ 10 cm);
- Serum immunoglobulin (IgM, IgA and IgG);
- β2 microglobulin (quantitative and categorical: ≤ 3 mg/L vs > 3 mg/L);
- Hepatitis B core antibody and hepatitis C virus antibody statuses (positive vs negative);
- Coagulation parameters of prothrombin time, international normalized ratio, and aPTT (< LLN vs normal vs > ULN);
- Mutation statuses: del 17p, TP53, trisomy 12, del 11q, del 13q, IGHV;
- Splenomegaly (yes vs no);
- Hepatomegaly (yes vs no);
- Bone marrow involvement (< 30% vs ≥ 30%);
- ALC;
- Hemoglobin (quantitative and categorical: ≤ 110 g/L vs. > 110 g/L);
- Platelet (quantitative and categorical: ≤ 100 x 10^9/L vs. > 100 x 10^9/L);
- ANC (quantitative and categorical: ≤ 1.5 x 10^9/L vs. > 1.5 x 10^9/L);
- Any cytopenia (yes: hemoglobin ≤ 110 g/L or platelet count ≤ 100 x 10^9/L or ANC ≤ 1.5 x 10^9/L vs no);
- LDH;
- Prior radiotherapy and prior transplants (yes vs no);
- Disease-related constitutional symptoms.

Categories of baseline characteristics can be modified depending on 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 patients reporting a history of any medical condition, as recorded on the CRF, will be summarized by system organ class (SOC) preferred term (PT), and treatment arm. A listing of medical history will be provided.

6.3.7 Prior Systemic Anti-Cancer Therapies

Number of prior lines of systemic anti-cancer 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 treatment arm using descriptive statistics. The therapies with the same sequence/line number are counted as one prior line of therapy. A listing of these therapies will be provided.

6.3.8 Prior and Concomitant Medications, Procedures and Surgeries

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 drug and were continuing at the time of the first dose of the study drug, or (2) started on or after the first dose date of the study drug up to 30 days after the last dose date or the day prior to initiation of a new CLL/SLL therapy, whichever occurs first.

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

In addition, the number and percentage of patients reporting postbaseline anti-cancer therapies will be summarized by WHO DD preferred name, and treatment arm. The number and percentage of patients reporting postbaseline cancer-related surgery or procedures will be summarized by treatment arm. A listing of procedures and surgeries will be provided.

6.4 EFFICACY ANALYSES

Statistical testing will be performed to compare the efficacy of zanubrutinib (Arm A) and ibrutinib (Arm B). For all efficacy analyses, patients will be analyzed according to the treatment arm to which they were randomized, and unless otherwise specified, efficacy analyses will be performed on the ITT Analysis Set. At the interim analysis of ORR, ORR, DOR and associated descriptive analyses will be performed on the first 415 randomized patients in the ITT Analysis Set.

Dates for investigator disease assessments are derived according to Appendix B, and best overall response derivation rules for CLL are derived according to Appendix C.

ORR, rate of PR-L or higher and rate of CR/CRi will be based on the best overall response (BOR), and BOR per investigator assessment and per independent central review will be subject to confirmation for CLL patients as detailed in Appendix C and in the separate independent review charter. BOR is defined as the best response from the randomization date to the data cutoff date, disease progression or the start of new CLL/SLL therapy, whichever comes first. Patients without any postbaseline disease assessment (regardless of the reason) will be considered as non-responders.

Stratified analyses will be based on the randomization stratification factors per IRT unless otherwise specified.

6.4.1 Primary Efficacy Analyses

The primary hypothesis testing for the primary endpoint of ORR per investigator assessment will be to demonstrate the noninferiority of zanubrutinib to ibrutinib.

Noninferiority testing for ORR

The null and alternative hypotheses for the noninferiority test are as follows:

- H_{0NI}: Response ratio (zanubrutinib/ibrutinib) ≤ 0.8558
- H_{aNI}: Response ratio (zanubrutinib/ibrutinib) > 0.8558

One interim analysis of ORR will occur approximately 12 months after 415 patients have been randomized, and the final analysis of ORR will occur approximately 12 months after 600 patients have been randomized.

The monitoring boundaries for the noninferiority test are based on the O'Brien Fleming boundary approximated by the Lan-DeMets spending function with an overall 1-sided level of 0.025. Hypothesis testing for the noninferiority of ORR at the interim analysis will be based on the first 415 randomized patients only and will have a 1-sided significance level of 0.005. Hypothesis testing for the noninferiority of ORR at the final analysis will be based on the entire ITT Analysis Set and will have a 1-sided significance level based on the actual information fraction (or covariance) of the interim and final test statistics. With 652 patients in the ITT Analysis Set at the final analysis, the actual information fraction is 64% (415/652), and the 1-sided significance level for the final analysis will be 0.0235.

The noninferiority hypothesis of ORR will be tested at each analysis using a stratified Wald test against a null response ratio of 0.8558, and the noninferiority of zanubrutinib to ibrutinib in ORR will be considered statistically significant if the 1-sided p-value is less than the 1-sided significance level at either the interim or the final analysis of ORR. The form of the Wald statistic (Z) will be as follows:

$$Z = \frac{\log(\rho) - \log(0.8558)}{\mathrm{SE}(\log(\rho))}$$

where log(ρ) is the log of the stratified Mantel-Haenszel response ratio estimate and SE(log(ρ)) is its corresponding standard error estimate (Greenland and Robins 1985)

Table 1 summarizes the interim and final analysis monitoring boundaries for ORR noninferiority and superiority testing with 64% information fraction.

Table 1. ORR Monitoring Boundaries

Analysis Timepoint	**ORR Noninferiority**		**ORR Superiority**	
	1-sided p-value boundary	**Approximate response ratio boundary [1]**	**1-sided p-value boundary**	**Approximate response ratio boundary [1]**
Interim Analysis of ORR	0.005	1.01	0.005	1.176
Final Analysis of ORR	0.0235	0.952	0.0235	1.105

[1] Approximate boundary based on assumed ibrutinib ORR of 70%.

Justification of the noninferiority margin for ORR

A non-inferiority margin of 0.8558 in response ratio was derived using the 95% to 95% fixed margin approach (FDA Guidance for Industry Non-Inferiority 2016). In the RESONATE trial (Byrd et al 2014), the ibrutinib effect over ofatumumab represented by the ratio of response rate (PR or higher) was 10.43 with a 95% CI of (5.2, 21.0) based on the independent review committee assessment. Thus, M1 is 5.2, the lower bound of the 95% CI. Since the effect size of ibrutinib is versus an active control (ofatumumab), rather than placebo, the choice of M1 is conservative, and a non-inferiority margin of 0.8558 (for the response ratio) retains over 90% of M1 (on the log scale).

Superiority testing for ORR

As described in Section 6.2.4, if the noninferiority in ORR per investigator assessment is statistically significant, then the superiority of zanubrutinib to ibrutinib in ORR will be tested. The null and alternative hypotheses for the superiority test are as follows:

- H_{0SUP}: Response ratio (zanubrutinib/ibrutinib) ≤ 1
- H_{aSUP}: Response ratio (zanubrutinib/ibrutinib) > 1

The monitoring boundaries for the superiority test are based on the O'Brien Fleming boundary approximated by the Lan-DeMets spending function with an overall 1-sided level of 0.025. If hypothesis testing for the superiority of ORR is performed at the interim analysis, it will be based on the first 415 randomized patients only and will have a 1-sided significance level of 0.005. If hypothesis testing for the superiority of ORR is performed at the final analysis, it will be based on the entire ITT Analysis Set and will have a 1-sided significance level that depends on the actual information fraction (or covariance) of the interim and final test statistics and will be the same 1-sided significance level used for the noninferiority ORR testing (0.0235, equivalent to a chi-squared p-value cutoff of 0.0469).

The superiority hypothesis of ORR will be tested using a stratified Cochran-Mantel-Haenszel test, and the superiority of zanubrutinib to ibrutinib in ORR will be considered statistically significant if the 1-sided p-value is less than the 1-sided significance level at either the interim or the final analysis of ORR.

Additional descriptive analyses

The 95% confidence interval (CI) for the response ratio will be constructed using a normal approximation. ORR will be summarized for each treatment arm along with its corresponding 95% CI.

Time to response, defined as the time from randomization to the earliest qualifying response, will be summarized descriptively for responders only.

Best overall response (BOR) will be summarized as well. In addition, the concordance between independent central review and investigator assessment of BOR will be summarized.

The rate of CR/CRi will be summarized for each treatment arm along with its corresponding 95% CI.

At the interim analysis of ORR only, the above descriptive analyses will be performed on the first 415 randomized patients in the ITT Analysis Set and BOR will also be summarized for the entire ITT Analysis Set.

6.4.2 Secondary Efficacy Analyses

The key secondary endpoints are PFS per investigator assessment and atrial fibrillation/flutter incidence. For the other secondary endpoints, the tests will be descriptive without multiplicity adjustment.

6.4.2.1 Key secondary endpoint: Progression-free Survival per investigator assessment

Hypothesis testing for the key secondary endpoint of PFS will first be to demonstrate the noninferiority of zanubrutinib to ibrutinib.

There will be a single analysis of PFS for the purpose of inference when approximately 205 PFS events have occurred; however, a 1-sided significance level of 0.00001 will be applied to each of the two descriptive analyses of PFS for the interim and final analyses of ORR to compensate for the potential type I error increase from the descriptive analysis. From the time of the ORR analyses to the analysis of PFS after 205 events have occurred, the sponsor will continue to maintain trial integrity according to the DIPP.

PFS will be right-censored for patients who meet one of the following conditions: 1) no baseline disease assessments; 2) PD or death more than 6 months from the last disease assessment (more than 12 months if a patient is on the disease assessment schedule of every 24 weeks); and 3) 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 2.

Table 2. 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/postbaseline disease assessments (and no death)	Date of randomization	Censored
Death or PD 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 patient is on the assessment schedule of every 24 weeks.

Noninferiority testing for PFS per Investigator Assessment

The null and alternative hypotheses for the noninferiority test are as follows:

- H_{0NI}: HR (zanubrutinib/ibrutinib) ≥ 1.3319
- H_{aNI}: HR (zanubrutinib/ibrutinib) < 1.3319

At the final analysis of PFS, hypothesis testing for the noninferiority of PFS per investigator assessment will be based on the entire ITT Analysis Set using a stratified Wald test and will have a 1-sided significance level of 0.02498. The form of the Wald statistic (Z) will be as follows:

$$Z = \frac{\beta - \log(1.3319)}{\mathrm{SE}(\beta)}$$

where β is the log of the hazard ratio estimate from the stratified Cox proportional hazards model and SE(β) is its corresponding standard error estimate

Justification of the noninferiority margin for PFS

A non-inferiority margin of 1.3319 was derived using the 95% to 95% fixed margin approach based on the RESONATE study. In the updated RESONATE results (Brown et al 2014), the estimated PFS HR for ibrutinib versus ofatumumab was 0.106 with a 95% CI of (0.073, 0.153). Therefore, the control arm effect (M1) is 0.153 in HR and -1.877 in log HR. A noninferiority margin of 1.3319 for the HR (zanubrutinib/ibrutinib) retains approximately 85% of M1 (on the log scale).

Superiority testing for PFS per Investigator Assessment

If the noninferiority of zanubrutinib to ibrutinib in PFS per investigator assessment is statistically significant, then the superiority in PFS per investigator assessment will be tested. The null and alternative hypotheses for the superiority test are as follows:

- H_{0SUP}: HR (zanubrutinib/ibrutinib) ≥ 1
- H_{aSUP}: HR (zanubrutinib/ibrutinib) < 1

Hypothesis testing for the superiority of PFS per investigator assessment will be based on the entire ITT Analysis Set using a stratified log-rank test and will have the same 1-sided significance level of 0.02498 (equivalent to a chi-squared p-value cutoff of 0.04996) used for the noninferiority PFS testing.

Additional descriptive analyses

The HR for PFS and its 95% CI will be estimated from a stratified Cox regression model.

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

Kaplan-Meier curves for PFS will be presented for each treatment arm. A listing of PFS-related information, eg, the date of the progression or censoring and the corresponding reasons, will also be provided.

6.4.2.2 Key secondary endpoint: Atrial fibrillation/flutter incidence

As described in Section 6.2.4, if the noninferiority of zanubrutinib to ibrutinib in ORR is statistically significant, then the superiority of zanubrutinib to ibrutinib in the key secondary endpoint of atrial fibrillation/flutter will be tested but separately from the fixed sequence hierarchical testing that includes ORR and PFS. The interim analysis will be performed on the Safety Analysis Set restricted to the first 415 randomized patients and according to the actual treatment received. The final analysis will be performed on the Safety Analysis Set according to the actual treatment received.

The monitoring boundaries for the superiority test are based on the O'Brien Fleming boundary approximated by the Lan-DeMets spending function with an overall 1-sided level of 0.025. If hypothesis testing for the superiority of the rate of atrial fibrillation/flutter is performed at the interim analysis, it will have a 1-sided significance level of 0.005 (equivalent to a chi-squared p-value cutoff of 0.0099). If hypothesis testing for the superiority of the rate of atrial fibrillation/flutter is performed at the final analysis, it will have a 1-sided significance level of 0.0235 (equivalent to a chi-squared p-value cutoff of 0.0469).

Hypothesis testing on the rate of atrial fibrillation/flutter will be performed using an unstratified chi-squared test if the expected counts in the 2 x 2 contingency table (treatment arm by atrial fibrillation/flutter status) are at least 5 patients. If any expected count in the 2 x 2 contingency table is less than 5 patients, then hypothesis testing will be performed using Fisher's exact test.

6.4.2.3 Other secondary endpoints

6.4.2.3.1 ORR per Independent Central Review

ORR per independent central review will be summarized and analyzed using the same methods used for ORR per investigator assessment.

6.4.2.3.2 PFS per Independent Central Review

PFS per independent central review will be summarized and analyzed using the same methods used for PFS per investigator assessment.

6.4.2.3.3 Duration of Response (DOR) by Independent Central Review

DOR by independent central review will be analyzed only for patients in the ITT Analysis Set who achieved a response (PR or higher) per independent central review. At the interim analysis of ORR only, DOR by independent central review will be analyzed only for the first 415 randomized patients in the ITT Analysis Set who achieved a response (PR or higher) per independent central review.

The censoring conventions and the analyses of DOR by independent central review will be the same as those of PFS with the exception that treatment arm comparisons will not be performed, ie, no HR or corresponding CI or testing will be calculated or performed.

6.4.2.3.4 DOR by Investigator Assessment

DOR by investigator assessment will be analyzed only for patients in the ITT Analysis Set who achieved a response (PR or higher) per investigator assessment. At the interim analysis of ORR only, DOR by investigator assessment will be analyzed only for the first 415 randomized patients in the ITT Analysis Set who achieved a response (PR or higher) per investigator assessment.

Analyses of DOR will be repeated based on the investigator assessment with the same methods as those for DOR by independent central review.

6.4.2.3.5 Time to Treatment Failure

The HR for time to treatment failure and its 95% CI will be estimated from a stratified Cox regression model.

The distribution of time to treatment failure, including the median and other quartiles, and the on-treatment rate (ie, no treatment failure) at selected timepoints such as 12, 18 and 24 months, will be estimated using the Kaplan-Meier method for each treatment arm. The 95% CI for the median and the other quartiles of time to treatment failure as well as the on-treatment rate at the selected timepoints will be estimated as for PFS. Time to treatment failure will be censored at the data cutoff for patients who did not discontinue.

6.4.2.3.6 Rate of PR-L or Higher by Independent Central Review

The rate of PR-L or higher by independent central review will be summarized for each treatment arm along with its corresponding 95% CI. At the interim analysis of ORR only, the rate of PR-L or higher by independent central review will be analyzed only for the first 415 randomized patients in the ITT Analysis Set.

6.4.2.3.7 Rate of PR-L or Higher per Investigator Assessment

The rate of PR-L or higher per investigator assessment will be summarized for each treatment arm along with its corresponding 95% CI. At the interim analysis of ORR only, the rate of PR-L or higher per investigator assessment will be analyzed only for the first 415 randomized patients in the ITT Analysis Set.

6.4.2.3.8 Overall Survival

Patients who remained alive as of the data cutoff or discontinued study due to reasons other than death will be right censored at the date on which the patient was last known to be alive. The methods and analyses for overall survival will be the same as those described for PFS.

6.4.2.3.9 Health Related Quality of Life

The scoring of the EORTC QLQ-C30 and EQ-5D-5L will follow their corresponding manuals (Fayers et al. 2001; EuroQol Group 1990; Herdman et al 2011).

EORTC QLQ-C30

The scores of the EORTC QLQ-C30 questionnaire will be summarized at each assessment timepoint for each treatment arm. Summaries will include: scores and changes from baseline in the QLQ-C30 global health status/QoL (GHS/QoL) scale and the five functional scales (Physical Functioning, Role Functioning, 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).

Higher QLQ-C30 symptom scales and single item scores indicate worse HRQoL. Higher scores in the functional scales and GHS/QoL scale indicate better HRQoL.

In addition, the change from baseline in QLQ-C30 GHS/QoL scale scores will be compared between treatment arms using a linear mixed model for repeated measures (MMRM) at the predefined timepoints of Cycle 7 (24 weeks) and Cycle 13 (48 weeks). The model will include the repeated measurement (for Cycles 4, 7, 10 and 13) of the change from baseline in QLQ-C30

GHS/QoL score as the dependent variable, and the treatment arm by assessment timepoint interaction, and baseline score as a covariate with either an unstructured covariance matrix (Mallinckrodt 2018) or a simpler covariance structure, eg, heterogeneous Toeplitz. The difference between treatment arms in change from baseline score at Cycles 7 and 13 as well as their 95% CIs will be estimated, and the corresponding p-values will be presented.

EQ-5D-5L

The EQ-5D-5L comprises a descriptive system and an EQ Visual Analogue Scale (EQ VAS) with the 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 by treatment arm at each assessment timepoint. Descriptive statistics will also be provided by treatment arm at each assessment timepoint including the number and percentage of patients reporting each level of problem on each dimension of the EQ-5D-5L.

6.4.3 Sensitivity Analyses

Primary endpoint of ORR

The noninferiority of the primary endpoint of ORR will also be analyzed in the Per-protocol Analysis Set.

A sensitivity analysis of ORR using an alternative set of response confirmation rules (see Appendix C) will be performed that allows PR-L to be subsequently confirmed as a best overall response of PR.

To account for disease progression due to study drug interruption, ORR and BOR will be summarized based on all disease assessments through the data cutoff date, disease progression or the start of new CLL/SLL therapy, whichever comes first; however, disease progression that occurs within 6 weeks of a study drug interruption of at least 7 days will not be counted as disease progression for the purpose of this sensitivity analysis.

To account for the impact of COVID-19, ORR will be summarized for each treatment arm excluding patients who have died due to COVID-19.

Key secondary endpoint of PFS

The non-inferiority of the key secondary endpoint of PFS will also be analyzed in the Per-protocol Analysis Set.

Alternative censoring rules such as censoring for new CLL/SLL therapies will be applied as another sensitivity analysis of PFS.

To account for disease progression due to study drug interruption, PFS will also be summarized where disease progression that occurs within 6 weeks of a study drug interruption of at least 7 days will not be counted as disease progression for the purpose of this sensitivity analysis.

To account for the impact of COVID-19, PFS will be summarized for each treatment arm while additionally censoring deaths due to COVID-19.

6.4.4 Subgroup Analyses

The primary and selected secondary endpoints will be summarized, as appropriate, in the following subgroups (if collected in both IRT and in the clinical data, clinical data including EDC will be used to define subgroups):

- Age group (< 65 vs ≥ 65 years);
- Sex;
- Geographic region (Asia vs Australia/New Zealand vs Europe vs North America);
- Prior lines of therapy (1-3 vs > 3);
- Baseline ECOG performance status (0 vs ≥ 1);
- Baseline del17p/TP53 mutation status (present or absent);
- Bulky disease (yes: any target lesion longest diameter ≥ 5 cm vs no);
- Baseline β2 microglobulin (≤ 3 mg/L vs > 3 mg/L);
- Baseline IGHV mutation status (unmutated vs mutated);
- Disease stage (Binet stage of A/B and Ann Arbor stage I-II bulky vs Binet stage C and Ann Arbor stage III/IV) at study entry

When the sample size of a subgroup is too small, the subgroup may be omitted from the analyses or be combined with other subgroups if appropriate. Forest plots for the subgroup analyses will be provided with 95% CIs based on unstratified methods.

6.4.5 Exploratory Efficacy Analyses

Cox and/or logistic regression models, as well as descriptive comparisons, may be used to explore the association between the prognostic, predictive biomarkers, as well as MRD and the clinical outcomes.

MRD

The percentage of CLL cells amongst leukocytes will be provided in a listing.

Self-administered Activity and Quality of Life Questionnaire

The self-administered activity and quality-of-life questionnaire data consist of the following and may be provided in listings if appropriate:

- 10 questions dealing with symptoms, activity and general well-being
- 3 questions related to contact with physician, study drug compliance and mobile device usage
- 6-minute walk test

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, and coagulation), vital signs, physical exams, ECOG performance status and ECGs findings will also be used to assess safety. Descriptive statistics will be used to summarize the safety data by treatment arm.

For all safety analyses by treatment arm, patients will be analyzed according to the actual treatment received.

6.5.1 Extent of Exposure

The extent of exposure to study drug will be summarized descriptively by treatment arm as the number of cycles received (number and percentage of patients), treatment duration (months), cumulative total dose received per patient (g), actual dose intensity (mg/day) and relative dose intensity (%).

The number and percentage of patients with dose reductions, dose interruption, and drug discontinuation will be summarized with the respective reasons.

The actual dose intensity is defined as the actual cumulative dose (mg) taken based on the total dose per day divided by the treatment duration (days). The relative dose intensity is defined as the ratio of the actual dose intensity to the planned dose intensity as a percentage where the planned dose intensity is 320 mg/day for zanubrutinib and 420 mg/day for ibrutinib.

Patient data listings will be provided for all dosing records.

6.5.2 Adverse Events

The AE verbatim descriptions (terms as recorded by the investigator on the eCRF) will be classified into standardized medical terminology using Medical Dictionary for Regulatory Activities (MedDRA). AEs will be coded to the MedDRA (Version 20.0 or higher) lower level term closest to the verbatim term. The linked MedDRA preferred term (PT) and primary system organ class (SOC) will also be captured in the database.

A treatment-emergent adverse event (TEAE) is defined as an AE that has an onset date on or after the first dose of study drug up to 30 days after the last dose of study drug or the day prior to initiation of a new CLL/SLL therapy, whichever occurs first. If a TEAE worsens to grade 5 more than 30 days after last dose of study drug and prior to initiation of a new CLL/SLL therapy, the grade 5 AE will be treatment-emergent. Only TEAEs will be included in summary tables. All AEs, treatment-emergent or otherwise, will be presented in patient data listings.

TEAEs will be summarized by treatment arm as the number and percentage of patients reporting TEAEs by SOC and PT. A patient will be counted only once within a system organ class and preferred term. TEAEs will also be summarized by grade and by the relationship to study drug assessed by the investigator, ie, treatment-related or not. A treatment-related AE is an AE that is assessed by the investigator as related to the study drug or is missing an assessment of the causal relationship. TEAEs summarized by grade or by relationship to study drug will be summarized at the maximum severity grade or strongest causal relationship to study drug, respectively.

No imputation of AE grades will be performed. Treatment-emergent adverse events (TEAEs) with missing CTCAE grade will only be summarized in the total column.

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

- TEAE;
- treatment-related TEAE;
- grade 3 or higher TEAE;
- treatment-related grade 3 or higher TEAE;
- TEAE leading to death;
- treatment-related TEAE leading to death;
- serious TEAE;
- treatment-related serious TEAE;
- TEAE leading to treatment discontinuation;
- treatment-related TEAE leading to treatment discontinuation;
- TEAE leading to dose modification;
- treatment-related TEAE leading to dose modification;
- TEAE leading to dose reduction;
- treatment-related TEAE leading to dose reduction;
- 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 grade.

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

- treatment-emergent AESIs;
- grade 3 or higher treatment-emergent AESIs;
- serious treatment-emergent AESIs;
- treatment-emergent AESIs leading to treatment discontinuation;
- treatment-emergent AESIs leading to dose modification;

- treatment-emergent AESIs leading to dose reduction;
- treatment-emergent AESIs leading to dose interruption;
- treatment-emergent AESIs by maximum severity.

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

TEAEs related to liver, arrythmia and diarrhea will be defined and summarized by category and PT for all such events, grade 3 or higher, serious and by maximum severity.

Incidence and time to diarrhea (grade 3 or higher), severe bleeding (defined as grade 3 or higher bleeding of any site or central nervous system bleeding of any grade), and atrial fibrillation (both new onset and exacerbation of existing atrial fibrillation) will also be summarized.

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. In addition, a listing of serious AEs for patients with COVID-19 infection will be provided. No imputation will be done in the AE listings.

6.5.3 Laboratory Values

The values and changes from baseline for selected CBC components, serum chemistry and serum immunoglobulin parameters will be summarized by treatment arm at each assessment timepoint and for the worst postbaseline assessment.

Certain laboratory parameters can come from central or local laboratories (eg, hematology, chemistry). Summaries of laboratory parameters will be based on the preferred laboratory (central or local, not both) for each patient, and the preferred laboratory will be the central laboratory if central laboratory data are available and the local laboratory if central laboratory data are not available.

Laboratory parameters that are graded in NCI-CTCAE v4.03 will be summarized by the CTCAE grade and laboratory parameters that are graded per Hallek (2008) including platelets, hemoglobin and absolute neutrophil count will be summarized by the Hallek toxicity grade. In the summary of laboratory parameters by CTCAE grade, parameters with CTCAE grading in both high and low directions (eg, calcium, glucose, magnesium, phosphorus, potassium, sodium) will be summarized separately.

A summary of the number and percentage of patients with grade 3 or higher toxicity will be provided for selected laboratory parameters of interest. Shift tables assessing the toxicity grade at baseline versus the worst postbaseline toxicity grade recorded will be presented.

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

Incidence of patients who met one or more of the Hy's law criteria will be summarized. A listing of patients 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 assessment timepoint. A listing by patient and assessment timepoint will be generated.

6.5.5 Electrocardiograms (ECG)

The QTc-Fridericia (QTcF) interval values and changes from baseline will be summarized by treatment arm for each assessment timepoint. If triplicate readings (eg, screening) are recorded, the average of the readings for the assessment timepoint will be used for the summary.

The number and percentage of patients satisfying the following QTcF conditions at any time postbaseline will be summarized:

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

A listing of abnormal ECG by patient and assessment timepoint will be generated.

6.5.6 ECOG Performance Status

ECOG performance status will be summarized by treatment arm at each assessment timepoint.

6.6 Pharmacokinetic Analyses

A population PK analysis may be performed to include plasma concentrations of zanubrutinib from this trial in an existing model. PK parameters such as apparent systemic clearance and AUC may be derived from the population PK analysis if supported by data.

An exposure-response (efficacy or safety endpoints) analysis may be performed if supported by data. The results from the population PK and exposure-response analyses may be reported separately from the Clinical Study Report.

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

7 INTERIM ANALYSIS

There will be one interim analysis for the noninferiority (and superiority if noninferiority is met) testing of ORR. The interim analysis will be performed approximately 12 months after the randomization of 415 patients. The monitoring boundaries for the interim and the final analyses for the nonferiority and superiority tests are based on the O'Brien-Fleming boundary approximated by the Lan-DeMets spending functions.

8 CHANGES IN THE PLANNED ANALYSIS

Not applicable.

9 REFERENCES

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

Byrd JC, Brown JR, O'Brien S, et al. Ibrutinib versus ofatumumab in previously treated chronic lymphoid leukemia. N Engl J Med. 2014;371(3):213-23.

Byrd JC, Hillmen P, O'Brien S, et al. Long-term follow-up of the RESONATE phase 3 trial of ibrutinib vs ofatumumab. Blood. 2019;133(19):2031-42.

Cheson BD, Byrd JC, Rai KR, et al. Novel targeted agents and the need to refine clinical endpoints in chronic lymphocytic leukemia. J Clin Oncol. 2012;30:2820-2.

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)

Greenland, S. and Robins, J. M. (1985), Estimation of a Common Effect Parameter from Sparse Follow-Up Data. Biometrics, 41, 55–68.

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

Hallek M, Cheson BD, Catovsky D, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukaemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute - Working Group 1996 guidelines. Blood. 2008;111(12):5446-56.

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.

Mallinckrodt, 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.

Maurer, W and Bretz, B. Multiple testing in group sequential trials using graphic approaches. Statistics in Biopharmaceutical Research. 2013; 5:4, 311-320.

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

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.

10. APPENDIX

APPENDIX A: IMPUTATION OF MISSING OR PARTIALLY MISSING DATES

In general, missing or partial dates will be imputed at the ADaM dataset level. The following rules will apply to the specific analysis and summary purposes mentioned below only. Date of death, discontinuation from study and analysis data cutoff will be accounted for in the date imputations as applicable.

A.1 Prior/Concomitant Medications/Procedures/Surgeries

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

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 month and day will be January 01 or the date of first dose of study drug if they have the same year
- If only day is missing, then the imputed day will be the first day of the month or the date of first dose of study drug if they have the same month and year

- If the start date is completely missing, the imputed date will be the date of first dose of study drug as long as the AE end date is not before the date of first dose of study drug.

If the 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 the end date is completely missing, do not impute.

A.3 Deaths

For partial death dates, impute as follows:

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

A.4 New Anti-cancer therapy

For partial dates of new anti-cancer therapies, dates will be imputed as for concomitant medication start dates.

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.

Appendix B: Derivation of Investigator Disease Assessment Dates

Dates corresponding to each investigator disease assessment (eg, CR, PR, PR-L) will be derived for disease assessments other than PD. Investigator assessments of PD will use the PD date provided directly by the investigator in the eCRF.

The investigator disease assessment is based on several individual assessments, eg, target and non-target lesions, liver/spleen and CBCs, and the disease assessment date derivation will be computed as the latest of the applicable individual assessments, which differ by investigator disease assessment and disease type (ie, CLL vs SLL) and are described in Table 3.

Table 3. Assessment Dates used for Investigator Disease Assessment Date Derivation

Assessment	Investigator Disease Assessment		
	CR, CRi	nPR	PR, PR-L, SD, NE
Target and non-target lesion imaging	Y	Y	Y
Liver / spleen imaging	Y	Y	Y
Physical exam of liver/spleen, lymph nodes	Y	Y	Y
CBC collection (CLL only)	Y	Y	Y
Bone marrow biopsy	N [1]	N [1]	N
Constitutional symptoms check (CLL only)	Y	Y	N

Y = assessment and corresponding date accounted for when determining investigator disease assessment date for given investigator disease assessment (eg, CR, PR); N = assessment not used for given investigator disease assessment. The latest of all dates with "Y" for the given investigator disease assessment will be the derived disease assessment date.

[1] bone marrow biopsy is relevant for certain disease assessments but will not be used to derive the assessment date.

APPENDIX C: DERIVATION OF BEST OVERALL RESPONSE FOR CLL

For CLL patients, the best overall response will require two assessments at least 8 weeks apart to confirm a best overall response of PR-L, PR, nPR, CRi or CR that meet the following criteria:

- To confirm a response of CR, there must be one CR and another CR or CRi in any order and can only have intervening responses of non-PD or NE and at most one intervening PR or nPR;
- To confirm a response of CRi, there must be two CRis and can only have intervening responses of non-PD or NE and at most one intervening PR or nPR;
- To confirm a response of nPR, confirmed CR or CRi must not be met and there must be EITHER an nPR with a subsequent nPR or higher OR an nPR or higher with a subsequent nPR, and can have intervening responses of PR-L, PR, non-PD, NE or SD;
- To confirm a response of PR, confirmed CR, CRi or nPR must not be met and there must be a PR or higher with a subsequent PR or higher and can have intervening responses of PR-L, non-PD, NE or SD;
- To confirm a response of PR-L, confirmed CR, CRi, nPR or PR must not be met and there must be a PR-L with a subsequent PR-L or higher or a PR-L or higher with a subsequent PR-L and can have intervening responses of non-PD, NE or SD;
- No other best overall response category requires confirmation.

Time to response will be the first PR or higher that is subsequently confirmed as a PR or higher.

For a sensitivity analysis of ORR with an alternative response confirmation, the best overall response will require two assessments at least 8 weeks apart to confirm a best overall response of PR-L, PR, nPR, CRi or CR that meet the following criteria:

- To confirm a response of CR, there must be one CR and another CR or CRi in any order and can only have intervening responses of non-PD or NE and at most one intervening PR or nPR;
- To confirm a response of CRi, there must be two CRis and can only have intervening responses of non-PD or NE and at most one intervening PR or nPR;
- To confirm a response of nPR, confirmed CR or CRi must not be met and there must be EITHER an nPR with a subsequent nPR or higher OR an nPR or higher with a subsequent nPR, and can have intervening responses of PR-L, PR, non-PD, NE or SD;

- To confirm a response of PR, confirmed CR, CRi or nPR must not be met and there must be a PR-L or higher with a subsequent PR or higher and can have intervening responses of PR-L, non-PD, NE or SD;
- To confirm a response of PR-L, confirmed CR, CRi, nPR or PR must not be met and there must be a PR-L or higher with a subsequent PR-L and can have intervening responses of non-PD, NE or SD;
- No other best overall response category requires confirmation.

For this sensitivity analysis, if a PR or higher is subsequently confirmed as PR or higher, then the date of first response will be the first PR or higher that is subsequently confirmed as a PR or higher, and if a PR-L is subsequently confirmed as PR, the date of first response will be the last PR-L or higher that is subsequently confirmed as a PR. If both scenarios apply, the date of first response will be the earlier of the two.

APPENDIX D: EORTC QLQ-C30 SCORING

The principle for scoring the QLQ-C30 applies to all scales/scores: raw scores are calculated as the average of the items that contribute to the scale.

A linear transformation to standardize the raw scores is utilized, so that the scores are ranged from 0 to 100. Increases in scores for functional domains (e.g., physical, role, social, emotional, etc.) are improvements while increases in scores for symptoms (e.g., fatigue, vomiting and nausea, diarrhea, pain, etc.) are deteriorations.

Missing Items

If at least half of the items for a scale are answered, then all the completed items are used to calculate the score. Otherwise, the scale score is set to missing.

In practical terms, if items I_1, I_2, ... I_n are included in a scale, the procedure is as follows:

Raw Score

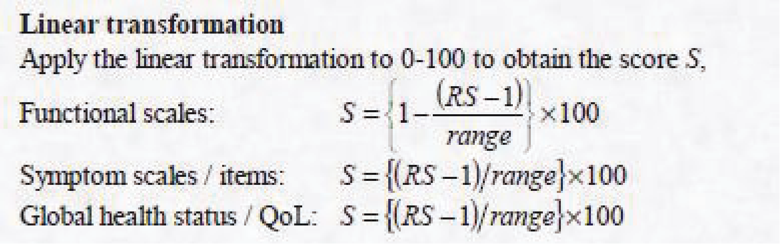
For all scores, the raw score (RS), is the mean of the component items

RS = $(I_1+I_2+\ldots+I_n)/n$

Derived Scale

The derived scales are obtained from the raw scores as defined in the EORTC manual. The derived scales have a more intuitive interpretation: larger function scale or global health status / QoL are improvements while larger symptom scales (e.g., pain, nausea, etc.) are deteriorations.

The derivation formulas are as follows:



Linear transformation

Apply the linear transformation to 0-100 to obtain the score S,

Functional scales: $S = \left\{1 - \frac{(RS-1)}{range}\right\} \times 100$

Symptom scales / items: $S = \{(RS-1)/range\} \times 100$

Global health status / QoL: $S = \{(RS-1)/range\} \times 100$

Scales of QLQ-C30

	Scale	Number of items	Item range	Item Numbers
Global health status/ QoL Global health status/QOL	QL2	2	6	29,30
Functional Scales				
Physical functioning	PF2	5	3	1, 2, 3, 4, 5
Role functioning	RF2	2	3	6, 7
Emotional functioning	EF	4	3	21, 22, 23, 24
Cognitive functioning	CF	2	3	20, 25
Social functioning	SF	2	3	26, 27
Symptom Scales/ items				
Fatigue	FA	3	3	10, 12, 18

Nausea and vomiting	NV	2	3	14, 15
Pain	PA	2	3	9, 19
Dyspnoea	DY	1	3	8
Insomnia	SL	1	3	11
Appetite loss	AP	1	3	13
Constipation	CO	1	3	16
Diarrhoea	DI	1	3	17
Financial Difficulties	FI	1	3	28

EQ-5D-5L: for the 5 level Dimensions, scores on a 5-point Likert scale of 1 to 5, with level 1 indicating no problem and level 5 indicating extreme problems. The Health State is defined by combining one level from each dimension. Lower scores indicate better health.

VAS includes a scale of 0 to 100 with higher scores indicating better health status.

Scales of EQ-5D-5L

MOBILITY
I have no problems in walking about □
I have slight problems in walking about □
I have moderate problems in walking about □
I have severe problems in walking about □
I am unable to walk about □

SELF-CARE
I have no problems washing or dressing myself □
I have slight problems washing or dressing myself □
I have moderate problems washing or dressing myself □
I have severe problems washing or dressing myself □
I am unable to wash or dress myself □

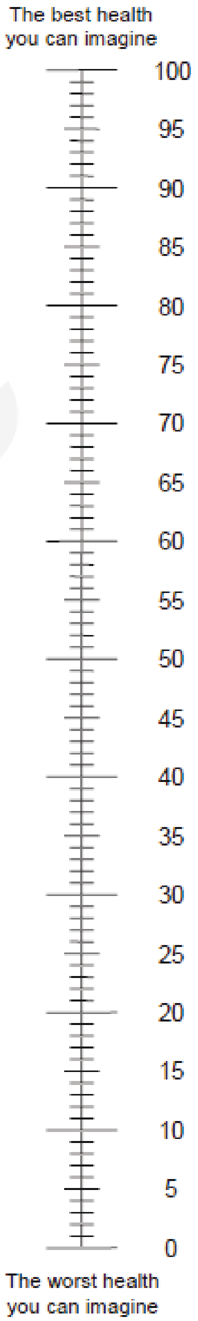
USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activities)
I have no problems doing my usual activities □
I have slight problems doing my usual activities □
I have moderate problems doing my usual activities □
I have severe problems doing my usual activities □
I am unable to do my usual activities □

PAIN / DISCOMFORT
I have no pain or discomfort □
I have slight pain or discomfort □
I have moderate pain or discomfort □
I have severe pain or discomfort □
I have extreme pain or discomfort □

ANXIETY / DEPRESSION
I am not anxious or depressed □
I am slightly anxious or depressed □
I am moderately anxious or depressed □
I am severely anxious or depressed □
I am extremely anxious or depressed □

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY = □


The best health you can imagine
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
The worst health you can imagine