STATISTICAL ANALYSIS PLAN

Protocol IPI-145-07

A Phase 3 Study of IPI-145 versus Ofatumumab in Subjects with Relapsed or Refractory Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma

Protocol Version:	**Amendment 3**
Type of Analysis Plan:	**Final Analysis**
Version:	**4.0**
Date:	**9 May 2017**
Author:	**Laura Truskowski, MS**

STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

The undersigned has developed this statistical analysis plan (SAP):

Name/Title	Signature	Date
Laura Truskowski, MS Principal Biostatistician		10 May 2017

The undersigned have reviewed this SAP and approve it in its entirety:

Name/Title	Signature	Date
Barry Turnbull, PhD Biostatistical Consultant		10 MAY 2017
Hahop Youssoufian, MD, MSc Head of Hematology and Oncology		10 May 2017

DOCUMENT HISTORY

Version	Date	Author(s)	Brief Summary of Changes
1.0	02 Oct 2014	Shijie Tang	Original
2.0	24 Nov 2015	Shijie Tang	Clarified the timing of the 2nd interim analysis of overall survival; added biomarkers to baseline characteristics; other changes. Refer to Section 8.2 for Details.
3.0	13 Apr 2016	Shijie Tang	In Section 3.5.2, the estimated numbers of OS events at each of OS interim analyses and final analysis are provided. Lymph node response rate and hematological improvement rate are removed from the hierarchy of inferential testing in secondary endpoints.
4.0	9 May 2017	Laura Truskowski	Update to Verstem as Sponsor, updates due to protocol amendment 3 including removal of 3 year limit on treatment duration.

Table of Contents

LIST OF ABBREIVATIONS

Abbreviation	Description
ADaM	Analysis Data Model
AE	Adverse Event
ANCOVA	Analysis of Covariance
AT	All-Treated
ALT	Alanine Aminotransferase
ANC	Absolute Neutrophil Count
AST	Aspartate Aminotransferase
ATC	Anatomical Therapeutic Chemical
BID	Twice a day
BOR	Best Overall Response
BUN	Blood Urea Nitrogen
CI	Confidence Interval
CLL	Chronic Lymphocytic Leukemia
CMH	Cochran–Mantel–Haenszel
CO2	Bicarbonate
CR	Complete Response
CRi	Complete Response with Incomplete Marrow Recovery
CS	Abnormal and Clinically Significant
CTCAE	Common Terminology Criteria for Adverse Events
DOR	Duration of Response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form
eDISH	Evaluation of Drug Induced Serious Hepatotoxicity
EQ-5D	European Quality of Life-5 Dimensions
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue
FACT-G	The Functional Assessment of Cancer Therapy - General

Abbreviation	Description
HR	Hazard Ratio
IDMC	Independent Data Monitoring Committee
IPI-145	(S)-3-(1-(9H-purin-6-ylamino)ethyl)-8-chloro-2-phenylisoquinolin-1(2H)-one
IPCW	Inverse Probability of Censoring Weighted
IRC	Independent Review Committee
IRT	Interactive Response Technology
ITT	Intent-To-Treat
IV	Intravenous
IWCLL	International Workshop on Chronic Lymphocytic Leukemia
IWG	International Working Group
LC/MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LDH	Lactate Dehydrogenase
LFT	Liver Function Test
MedDRA	Medical Dictionary for Regulatory Activities
MEOI	Medical Events of interest
MMRM	Mixed Effect Repeated Measures
MRD	Minimal Residual Disease
NCI	National Cancer Institute
NCS	Abnormal but Not Clinically Significant
NED	No Evidence of Disease
ORR	Overall Response Rate
OS	Overall Survival
PD	Progressive Disease
PFS	Progression-Free Survival
PK	Pharmacokinetics
PP	Per-Protocol
PR	Partial Response
PRwL	PR with Lymphocytosis

Abbreviation	Description
PT	Preferred Term
QoL	Quality of Life
RBC	Red Blood Cell
RPSFT	Rank Preserving Structural Failure Time
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
SI	Standard International System of Units
SLL	Small Lymphocytic Lymphoma
SOC	System Organ Class
SPD	Sum of Products
TEAE	Treatment Emergent Adverse Event
TTR	Time to Response
UNK	Unknown
VAS	Visual Analog Scale
WHO	World Health Organization
WHO-DD	World Health Organization Drug Dictionary

1 INTRODUCTION

This is the statistical analysis plan (SAP) for study IPI-145-07, *A phase 3 study of IPI-145 versus ofatumumab in subjects with relapsed or refractory chronic lymphocytic leukemia/small lymphocytic lymphoma*. This SAP is prepared according to Amendment 3 of the protocol, dated 09 February 2017.

2 STUDY OBJECTIVES

2.1 Primary Objective

The primary objective of this trial is to examine the efficacy of duvelisib (IPI-145) monotherapy versus ofatumumab monotherapy in subjects with relapsed or refractory Chronic Lymphocytic Leukemia (CLL) or Small Lymphocytic Lymphoma (SLL).

2.2 Secondary Objectives

- To determine the safety of duvelisib in subjects with CLL or SLL
- To evaluate the pharmacokinetics (PK) of duvelisib and, if applicable, its metabolite(s)

2.3 Exploratory Objectives

- To evaluate the health-related quality of life (QoL) of subjects
- To evaluate pharmacodynamic biomarkers of duvelisib
- To evaluate biomarkers that may predict duvelisib clinical activity and/or safety
- To evaluate mechanisms of resistance in subjects who exhibit disease progression while being treated with duvelisib or ofatumumab
- To evaluate genomic features of tumors predictive of response in subjects treated with duvelisib or ofatumumab

3 STUDY DESIGN

3.1 Overview

Study IPI-145-07 is a two-arm, randomized (1:1 ratio), open-label, phase 3, superiority trial designed to evaluate the efficacy and safety of duvelisib compared to ofatumumab.

Subjects who meet all the eligibility criteria at Screening will return to the clinic on Day 1 to receive their first dose of study drug (randomized to either duvelisib or ofatumumab). The first treatment cycle for each treatment arm will be 3 weeks (21±2 days). Subsequent treatment cycles will be 4 weeks (28±4 days).

Subjects randomized to duvelisib will be given a starting dose of 25 mg duvelisib administered orally twice daily (BID) initially in a 21-day treatment cycle followed by 28-day treatment cycles for up to 18 cycles or until disease progression or unacceptable toxicity (whichever comes first). After completing approximately 18 cycles of treatment, subjects who, in the opinion of the investigator, may derive benefit from continued treatment may receive additional cycles of duvelisib until disease progression or unacceptable toxicity. Subjects randomized to ofatumumab will receive treatment consistent with the approved product labeling: 8 weekly infusions, starting with an initial IV dose of 300 mg ofatumumab on Day 1 followed by seven weekly doses of 2000 mg. Thereafter, subjects will receive 2000 mg ofatumumab once every month for four months or until disease progression or unacceptable toxicity (whichever comes first), as outlined in the schedule of assessments. Administration of ofatumumab will not exceed the 12 doses (within 7 cycles) as described in the prescribing information.

Subjects will be followed for survival for up to 6 years (or other duration as specified by the protocol) from randomization or until death. Follow-up visits will occur every 6 months which can be conducted through a telephone interview.

3.2 Sample Size Consideration

This study employs a randomized, open-label, parallel design to assess the potential superiority of duvelisib treatment over ofatumumab treatment on progression-free survival (PFS) in CLL or SLL subjects.

The sample size was determined to provide sufficient power to test the primary endpoint of PFS in a group sequential design with one planned interim analysis.

Assuming an exponential distribution for PFS, a total of 185 PFS events (deaths or progressions) will provide approximately 93% power to detect a hazard ratio of 0.6 (median PFS of 9 months in the ofatumumab group versus 15 months in the duvelisib group) using a one-sided log-rank test at a 2.5% overall significance level (ie, alpha of 0.025). The study design employs the Lan-DeMets spending function for O'Brien-Fleming boundary as the alpha spending function and the Hwang-Shih-DeCani gamma (-4) spending function as the beta spending function. A total of 300 subjects will be randomized in a 1:1 ratio to receive either ofatumumab or duvelisib. Assuming a peak accrual rate of 30 subjects per month with 11 months for accrual ramp-up and a 4% cumulative dropout rate per year, the enrollment would complete in 16 months, with the final

analysis for PFS estimated to occur approximately 27.5 months from the randomization of the first subject.

3.3 Randomization

Once a subject has met all entry criteria, the Interactive Response Technology (IRT) will be used to generate a distinct subject identifier. If a subject discontinues from the study, the subject identifier will not be reused and the subject will not be allowed to re-enter the study.

Eligible subjects will be randomized via IRT in a 1:1 ratio to one of two treatment arms:

- Arm 1: duvelisib 25 mg BID
- Arm 2: Ofatumumab

In order to ensure subject balance between treatment groups, study subjects will be stratified by the following:

- High risk cytogenetics (presence vs absence of del[17p])
- Refractory/early relapse to purine analog based treatment (progression <12 months after fludarabine/pentostatin: yes vs no)
- Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)

First dose is to occur within 7 days of randomization after all screening assessments have been completed.

3.4 Blinding

This study is open-label and Investigators, site staff, Sponsor and Sponsor designees will have access to treatment assignments for individual subjects. To reduce bias, a blinded independent central review of disease status will be conducted, which will be blinded to individual subject treatment assignments. Investigators, site staff, Sponsor and Sponsor designees will only have access to blinded (pooled) aggregate study data, except for a select number of Sponsor staff, who may review aggregate SAEs and Medical Events of Interest (MEOI) data by treatment groups.

3.5 Planned Analyses

Interim and final analyses are planned for PFS and OS. For an interim analysis, the actual p-value boundaries will be calculated based on the actual number of PFS and OS events at the analysis. For the final analysis, the actual p-value boundary for efficacy will be calculated based on the actual number of PFS and OS events at the interim analysis (or analyses) and the final analysis.

The Independent Data Monitoring Committee (IDMC) will review PFS, OS and other efficacy data at the time of PFS interim analysis. The list of IDMC deliverables, including tables and listings, is included in the IDMC charter. These deliverables will be based on the methods described in this SAP.

3.5.1 Planned Analyses for PFS

There will be two planned analyses for PFS, one interim analysis and one final analysis. After approximately 50% of the planned PFS events (ie, 93 PFS events) have been observed based on the blinded, independent, central review, an interim analysis of efficacy will be performed.

At the interim analysis, PFS will be tested at the alpha level based on the Lan-DeMets alpha spending function for the O'Brien-Fleming boundary with the opportunity to stop the study (other than survival follow-up) for overwhelming evidence of efficacy. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for efficacy is one-sided p-value of 0.0015 (corresponding approximately to a HR of 0.540). In the meantime, PFS will also be tested at the alpha level for futility based on the alpha and beta spending functions specified for the study design. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for futility is one-sided p-value of 0.4791 (corresponding approximately to a HR of 0.990). The futility boundary of this study is non-binding, meaning that the Type I error is properly controlled even if the study is continued after the futility boundary for PFS is crossed at the interim analysis. Under the protocol design assumptions (eg, the study is fully enrolled in 16 months), the interim analysis is expected to occur approximately 17 months after the first subject is randomized.

If the study is not stopped at the interim analysis, the final analysis will be performed when approximately 185 PFS events have occurred in the study. If the interim analysis and the final analysis occur at exactly 93 and 185 PFS events, the criterion of statistical significance (ie, boundary for efficacy) is one-sided p-value of 0.0245 (corresponding approximately to a HR of 0.748).

3.5.2 Planned Analyses for OS

There will be three planned analyses for OS: two interim analyses and one final analysis. The first interim analysis of OS will be performed at the time of the planned PFS interim analysis after 93 PFS events have occurred. The estimated number of OS events at this first OS interim analysis is 24. The second interim analysis of OS will be performed at the planned PFS final analysis, after 185 PFS events have occurred. If the planned final analysis of PFS is not performed due to an early stop for efficacy, then the second interim analysis of OS will be conducted approximately 1 year after the first interim OS analysis. The estimated number of OS events at the second OS interim analysis is 58. The final analysis of OS will take place after the completion of follow-up for all subjects. The estimated number of OS events at the final OS analysis is 161. The information fractions at the two interim OS analyses and the final OS analysis will be calculated based on 161 planned OS events at the final OS analysis. The cumulative alpha spending at the three analyses will be 0.0001, 0.0249, and 0.0250, respectively. If the OS analyses occur at 24, 58, and 161 OS events, the p-value stopping boundaries for the three analyses will be 0.0001, 0.0249 and 0.0003, respectively. The actual p-value stopping boundaries will be derived from the actual number of OS events that occurred. The statistical

significance of OS will be claimed based on the hierarchical testing procedure as specified in Section 6.3.

4 ANALYSIS SETS

4.1 Intent-To-Treat (ITT) Analysis Set

The intent-to-treat (ITT) analysis set includes all subjects who are randomized, with treatment group designated according to randomization.

This analysis set will serve as the primary analysis set for all efficacy endpoints and demographics.

4.2 All-Treated (AT) Analysis Set

The all-treated (AT) analysis set includes all subjects who receive any amount of study drug (duvelisib or ofatumumab), with treatment group designated according to actual study treatment received. The AT analysis set will be the primary analysis set for safety endpoints.

4.3 Per-Protocol (PP) Analysis Set

The per-protocol (PP) analysis set includes all subjects in the ITT analysis set who do not violate the terms of the protocol in a way that would significantly affect the study outcome, with treatment group designated according to randomization. Subjects who meet any of the following criteria may be excluded from this analysis set:

- Do not have documented diagnosis of CLL or SLL
- Do not have measurable nodal disease at baseline as determined by the IRC
- ECOG performance status >2
- History of Richter's transformation or prolymphocytic leukemia
- Refractory to ofatumumab (defined as progression or relapse <12 months of receiving ofatumumab monotherapy or <24 months of receiving an ofatumumab-containing regimen)
- Prior exposure to a PI3K inhibitor
- Receive concomitant prohibited anticancer therapy
- Permanent discontinuation from study drug due to non-compliance

The PP analysis set will be a secondary analysis set for selected efficacy analyses.

5 STUDY ENDPOINTS

5.1 Primary Endpoint

The primary efficacy endpoint PFS is defined as time from randomization to the first documentation of progressive disease (PD) as determined by independent review or death due to any cause.

The censoring method of the primary endpoint can be found in Appendix A. The detailed algorithm will be specified in the Analysis Data Model (ADaM) specifications.

5.2 Secondary Endpoints

The secondary efficacy endpoints of the study are:

- Overall Response Rate (ORR), with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR), or PR with lymphocytosis (PRwL), according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Overall Survival (OS), defined as time from randomization to death
- Lymph node response rate, with lymph node response defined as ≥50% decrease in the SPD of target lymph nodes
- Hematologic improvement rate, defined as any of following hematologic improvement sustained for at least 60 days without transfusion or exogenous growth factors:
    - Neutrophil count >1,500/µL or an increase ≥50% from Baseline; or
    - Hemoglobin >11 g/dL or an increase ≥50% from Baseline; or
    - Platelet count >100,000/µL or an increase ≥50% from Baseline
- Duration of Response (DOR), defined as time from the first documentation of response to first documentation of PD or death due to any cause.

Other secondary endpoints of the study are:

- TEAEs and changes in safety laboratory values
- PK parameters derived from plasma duvelisib concentrations and, if applicable, its metabolite(s)

5.3 Exploratory Endpoints

The exploratory efficacy endpoints of the study are:

- Improvement in disease-related symptoms of fatigue, weight loss, unexplained fevers and night sweats from Baseline
- Minimal Residual Disease (MRD) in subjects with documented CR or CRi
- Time to Response (TTR), defined as the time from randomization to the first documentation of response (CR, CRi, PR or PRwL)

The QoL endpoints of the study are:

- Health-related QoL of subjects as assessed by the following subject-reported questionnaires:
    - EuroQol-5D (EQ-5D)
    - Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F)

The biomarkers of the study are:

- Serum and tissue biomarkers and blood immunophenotype
- Tumor genomic features (eg, DNA sequence variation, DNA copy number variation, and/or RNA expression)
- Germline DNA sequence variations

6 GENERAL STATISTICAL METHODS AND DATA HANDLING

6.1 General Methods

Summary statistics will be presented by treatment group, unless stated otherwise.

Unless otherwise specified, descriptive statistics for continuous data will include the number of subjects with data to be summarized (n), mean, standard deviation, 25% quartile, median, 75% quartile, and minimum and maximum. The same number of decimal places as in the raw data will be presented when reporting the minimum and maximum, one more decimal place than the raw data will be presented when reporting mean and median, and 2 more decimal places than the raw data will be presented when reporting standard deviation.

Descriptive statistics for categorical/qualitative data will include frequency counts and percentages. The total number of subjects in the treatment group will be used as the denominator for percent calculations, unless stated otherwise. All percentages will be presented with one decimal, unless otherwise specified. Percentages equal to 100 will be presented as 100, and percentages will not be presented for zero frequencies.

Descriptive statistics associated with time-to-event analyses will include the number of events, the number of subjects censored, 25% quartile, median, 75% quartile, and 95% confidence interval for median. These statistics will be presented for all time-to-event analyses, unless stated otherwise.

Listings will be provided for selected endpoints.

6.2 Handling of Missing Data

In general, values for missing data will not be imputed unless methods for handling missing data are specified.

6.2.1 Handling of Missing Dates/Months/Years for Adverse Events

Adverse events (AEs) with incomplete onset dates will be handled as follows for the sole purpose of determining treatment emergence (TEAE is defined in Section 7.3.1):

- If the start/end date of an AE is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The AE will be assumed to be treatment emergent if it cannot be definitively shown that the AE did not occur or worsen during the treatment-emergent period (worst case approach). The detailed algorithm will be specified in ADaM specifications.
- If the start date is completely missing, an AE will be considered treatment-emergent unless the stop date is before study drug administration.
- If the dose start date is missing for a subject at a data-cut, all AEs of the subject will be considered treatment-emergent.

The original partial or missing date will be shown in listings of AEs.

6.2.2 Handling of Missing Dates/Months/Years for Concomitant Medications

Prior or concomitant medications with incomplete start dates will be handled as follows for the sole purpose of determining whether a non-study medication is a concomitant medication:

- If the start/stop date of a medication is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The medication will be assumed to be concomitant if it cannot be definitively shown that the stop date is before the start of administration of study drug, or the start date is more than 30 days after the last date of administration of study drug. The detailed algorithm will be specified in ADaM specifications.
- If the start/stop dates are completely missing, a medication will be considered concomitant.
- If the dose start date is missing for a subject at a data-cut, all non-study medications of the subject will be considered concomitant.

The original partial or missing date will be shown in listings of all non-study medications.

6.2.3 Handling of Missing Dates/Months/Years for Disease History and Prior Therapies

For the purpose of calculating the duration from initial diagnosis, most recent relapse/refractory diagnosis or most recent prior therapy to randomization, partial/missing dates for diagnosis and last prior therapy completion will be imputed as follows:

- If both date and month are missing and the year is prior to the year of screening, the imputed date and month will be 01 July.
- If both date and month are missing and the year is the same as the year of screening, the imputed date will be the middle point between 01 Jan of the year and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- If date is missing and the month and year are prior to the month and year of screening, the imputed date will be 15th day of the month.
- If date is missing and the month and year are the same as the month and year of screening, the imputed date will be the middle point between the first date of the month and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- No imputation will be performed if the year is missing.

6.3 Multiple Comparisons/Multiplicity Adjustment

Of the secondary endpoints, ORR and OS are designated as key secondary efficacy endpoints. The primary endpoint and key secondary endpoints will be tested at an overall one-sided alpha level of 0.025 based on a gatekeeping approach.

If the primary endpoint is significant, the two key secondary endpoints will be sequentially tested in the order listed above. ORR will be tested at the one-sided 0.025 level only if PFS is declared statistically significant. OS will be tested according to the planned alpha spending specified in Section 3.5.2 only if PFS and ORR are declared statistically significant. If any null hypothesis is not rejected in this sequence of tests, formal sequential testing will be stopped.

6.4 Adjustments for Covariates

Adjustments for covariates will be considered for analysis of primary and key secondary endpoints, with details provided in Sections 7.2.1 and 7.2.2.

6.5 Subgroups

PFS, ORR and OS will be examined in the subgroups based on the following baseline characteristics or variables:

- Stratification factors (1. captured from central lab; 2 and 3 captured on the eCRF):
    1 High-risk cytogenetics (presence vs absence of del[17p])
    2 Refractory/early relapse to purine analog-based therapy (defined as progression <12 months after fludarabine/pentostatin: yes vs no)
    3 Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)
- Diagnosis (CLL or SLL)
- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Previously treated with ofatumumab (yes vs no, if sample size allows the analysis)
- Time from last dose of most recent prior anti-cancer therapy to randomization <12 months: yes vs no
- del[17p] or TP53 mutation (either or both present vs neither present; del[17p] will be as captured from central lab)

TEAEs (All Causalities) and TEAEs (Treatment-Related) will be examined in the following subgroups:

- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Baseline ECOG performance status (0 or 1 vs 2)

More details will be specified in Sections 7.2.1, 7.2.2 and 7.3.1.

6.6 Visit Windows

All data will be categorized based on the scheduled visit at which it was collected. These visit designators are predefined values that appear as part of the visit tab in the eCRF. There will be no additional analysis windowing done based on the assessment date.

6.7 Unscheduled Visits

Unscheduled visits will not be included in by-visit summary tables, unless otherwise specified. For laboratory tests, data from unscheduled visits will be included in listings and summaries of maximum changes from baseline, and the best or worst post-baseline values. For endpoints

based on disease status assessment, data from unscheduled assessments will be included in the derivation and analyses of the endpoints.

6.8 Baseline Values

Unless otherwise specified, the baseline value is defined as the value collected at the time closest to, and prior to, the start of study drug administration. Values collected at unscheduled visits prior to the start of the study drug administration will be included in the calculation of baseline values.

6.9 Computing and Coding Standards

Activities will be performed using the following tools:

Table, listing, and figure production	SAS Version 9.3 or higher
Coding	
Adverse Events	MedDRA Version 16.1 or higher
Medical Histories	MedDRA Version 16.1 or higher
Prior and Concomitant Medications	WHODrug Version September 2013
Grading	
AEs	CTCAE Version 4.03
Labs	CTCAE Version 4.03

7 STATISTICAL ANALYSES

7.1 Study Subjects

7.1.1 Disposition of Subjects

The disposition of subjects will include the number and percentage of subjects for the following categories: the number randomized, the number and percentage randomized but not dosed, the number and percentage dosed. These categories will be summarized for each treatment arm and for the two treatment arms combined (total). The percentages will be based on all randomized subjects (ITT analysis set).

An end-of-treatment disposition (still on treatment vs discontinued from treatment) will be provided for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set. The primary reason for treatment discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

An end-of-study disposition (still on study vs discontinued from study) will also be provided for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The primary reason for study discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

A summary of strata as captured in IRT will be presented by treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. A summary table and a listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF will be presented.

7.1.2 Protocol Deviations

Protocol deviations will be categorized as major or minor prior to data release for the interim or final analysis of the primary endpoint. A summary table of the major protocol deviations will be provided by treatment arm and for two arms combined (total) for the ITT analysis set. A listing of major protocol deviations will be provided, broken down by center and treatment arm.

7.1.3 Demographic and Other Baseline Characteristics

Demographic and other baseline variables will be summarized for each treatment arm based on all randomized subjects (ITT analysis set) and for the two treatment arms combined (total). The variables will include age, age group (<65 versus >=65), sex, race, ethnicity, height, weight and the following biomarkers:

- del[17p] (captured from central lab, presence vs absence vs indeterminate)
- del[17p] (captured from local lab, presence vs absence vs indeterminate)
- IgHV status (mutated vs unmutated vs indeterminate)
- CD38 (Positive (≥30%) vs Negative (<30%) vs indeterminate)
- ZAP70 (Positive (>19%) vs Negative (≤19%) vs indeterminate)
- del[17p] and/or TP53 mutation (either or both present vs neither present vs indeterminate; del[17p] will be as captured from central lab)

- TP53 mutation (presence vs absence vs indeterminate)

A separate table will be provided to compare del[17p] values captured from central lab and local lab. Demographic and other baseline variables will also be summarized based on the PP analysis set.

7.1.4 Disease History

Disease history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include diagnosis (CLL or SLL), risk factors used for randomization that are reported on eCRF (high risk cytogenetics [17p deletion: presence vs absence], refractory/early relapse to purine analog based treatment [yes vs no], grade 4 cytopenia(s) [presence vs absence]), years from initial diagnosis to randomization, months from most recent relapse/refractory diagnosis to randomization, Binet/Rai stage at initial diagnosis, type of prior treatment, Binet/Rai stage at baseline, and baseline lymphocyte count.

The durations to be summarized are defined as follows:

- Years from initial diagnosis to randomization will be calculated as (date of randomization – date of initial diagnosis + 1)/365.25.
- Months from most recent relapse/refractory diagnosis to randomization will be calculated as (date of randomization – date of most recent relapse/refractory diagnosis + 1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.5 Prior Therapies

Prior therapies will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include number of prior systemic therapies (summarized as a continuous variable and as a categorical variable) and months from most recent prior systemic therapy to randomization.

The duration to be summarized is defined as follows.

- Months from most recent prior systemic therapy to randomization will be calculated as (date of randomization – stop date of most recent systemic therapy +1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.6 Medical History

Medical history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set.

Medical history will be summarized by system organ class (SOC) and preferred term (PT) using the number and percentage of subjects who had at least one occurrence of an SOC or PT. The summary will be sorted alphabetically in SOC and by decreasing frequency of PT in the duvelisib arm within an SOC.

7.1.7 Prior and Concomitant Medications

Medications will be considered as prior if they stopped before the date of first dose of study drug.

Medications will be considered concomitant if they were taken at any time between the date of first dose of study drug and 30 days after the date of last dose of study drug, inclusive. If the start date or end date of a medication is completely or partially missing, refer to Section 6.2.2 for the algorithm to determine whether a medication is concomitant.

Prior medications and concomitant medications will be summarized separately. Both summaries will be based on the AT analysis set.

Medications will be summarized by ATC level 1, ATC level 2, and preferred drug name for each treatment arm. The summary will be sorted by decreasing frequency in ATC level 1, ATC level 2 and preferred drug name in the duvelisib arm. A subject taking the same drug multiple times will only be counted once.

A listing will be provided for all non-study medications taken on the study. An identifier will be provided to show if a medication is prior or concomitant. Medications that started more than 30 days after the last dose of study drug will be identified as post-treatment.

7.1.8 Exposure to Study Drug

Extent of exposure will be summarized for each treatment arm based on the AT analysis set.

Extent of exposure will be summarized for the following variables:

- Duration (weeks): (date of last dose – date of first dose + 1) divided by 7
- Number of cycles started (continuous and categorical)
- Ofatumumab arm only: For each infusion (1-12), number and percent of subjects receiving the infusion
- Relative dose intensity, defined as 100% x (total dose received)/ (planned cumulative dose for the duration of treatment)
- Number and percentage of subjects with a dose reduction
- duvelisib arm only: Number and percentage of subjects with a dose increase
- Number and percentage of subjects with a dose hold
- Number and percentage of subjects with study drug discontinued

A by-subject listing will be presented for exposure to study drug.

7.2 Efficacy Analyses

For PFS, ORR, lymph node response rate, DOR and other endpoints derived from progression and/or response status, the primary analyses will be based on the endpoints derived from

independent central review by Independent Review Committee (IRC). The endpoints derived from investigator assessment will be used in sensitivity analyses.

All efficacy analyses will be based on the ITT analysis set unless stated otherwise. If analyses are performed on more than one analysis set, the analyses on the ITT analysis set will be considered primary.

For stratified analysis of any efficacy endpoint, the following algorithm will be used to pool strata if there is insufficient information in any stratum (ie, there are <6 subjects, or there is no event for a time-to-event endpoint, or all subject have the same outcome for a binary endpoint in a stratum). (1) Strata will be ranked from the smallest to the largest based on the number of subjects, and if there is a tie, based on the number of events or responses. If there is still a tie, based on the reverse order of strata as determined by the stratification factors and levels (see Section 3.3). (2) The smallest stratum will be compared with the criteria of insufficient information. (3) If there is insufficient information in the smallest stratum, that stratum will be pooled with the smallest of the adjacent strata, which are defined as strata having 2 of the 3 stratification factors being at the same level as the smallest stratum. At the end of the three steps, the pooled stratum will replace the original two contributing strata. If there is still insufficient information in any stratum, the three steps will be repeated with the last pooled stratum assuming the stratum label of the larger of the two contributing strata for the purpose of additional pooling. The resulting pooled strata will be the strata used for the stratified analysis described in this section.

Listings of all efficacy endpoints will be provided.

7.2.1 Analyses of Primary Endpoint

7.2.1.1 Primary Analyses of PFS

The primary analyses of PFS will use PFS based on IRC assessment in the ITT analysis set.

Number of subjects with events, types of events (progression or death before progression), number of subjects censored, number of subjects for each reason of censoring, estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points may also be presented.

A stratified log-rank test (one-sided) will be used to compare PFS of the duvelisib arm against PFS of the ofatumumab arm at the interim and final analyses with the overall one-sided significance level controlled at 0.025. The nominal p-value required for statistical significance at an analysis will be calculated based on the numbers of PFS events at all preceding and current analyses. The HR (duvelisib/ofatumumab) and the corresponding 2-sided 95% CI will be estimated using a stratified Cox proportional hazards model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

PFS will be plotted for each treatment group using the Kaplan-Meier method. The plot will include the number of subjects at risk for each treatment group over time.

7.2.1.2 Sensitivity Analyses of PFS

The following sensitivity analyses will be performed:

- PFS based on investigator assessment: This endpoint will be analyzed using the same methods as the primary analyses described for PFS based on IRC assessment. In addition, the differences between investigator assessment and IRC assessment will be summarized.
- Worst-case sensitivity analysis: Subjects who are alive and have not had documented progression by data cutoff and who are "lost to follow-up" (missing at least one disease assessment right before data cut off) will be treated as censored at their last adequate disease assessment if they are on the control arm (ofatumumab) and treated as having a PFS event at the time of the next scheduled assessment following the last adequate disease assessment if they are on the experimental arm (duvelisib). PFS based on IRC assessment with the above worst-case censoring/event rule will be analyzed using the same methods as the primary analyses for PFS.
- Event-free survival (EFS): This is defined as time from randomization to the first documentation of PD as determined by IRC, start of new anticancer treatment or procedure, or death due to any cause. The event/censoring method is presented in Appendix B. EFS will be analyzed using the same methods as the primary analyses for PFS except that the summary of types of events will include a category of new anticancer treatment or procedure.
- Analysis using the PP analysis set.
- Analysis using the AT analysis set.
- Unstratified analyses: An unstratified log-rank test (one-sided) will be used to compare the two trreatment arms. An unstratified Cox regression (one-sided) will be used to estimate the hazard ratio (duvelisib/ofatumumab) with its 95% confidence interval. Other aspects of the analyses will be the same as the primary analyses of PFS.
- Cox regression with baseline covariates: A stratified Cox regression will be used to test treatment effect on PFS, adjusting for demographic and other baseline characteristics. The strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. A stepwise variable selection will be performed to choose the variables in the Cox regression. Candidate variables are age, gender, race, disease diagnosis (CLL or SLL), years from initial diagnosis, months from most recent relapse/refractory diagnosis, stage at diagnosis, stage at baseline, and number of prior systemic therapies. PFS based on IRC assessment will be used for this analysis.

7.2.1.3 Subgroup analyses of PFS

Subgroup analyses for PFS will be performed using the subgroups specified in Section 6.5. The HR (duvelisib/ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The hazard ratios and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

For selected subgroups, additional analyses of PFS will be performed. Number of subjects with events, types of events (progression or death before progression), number of subjects censored, number of subjects for each reason of censoring, estimates and 2-sided 95% confidence intervals for the 25th percentile, median, and 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points may also be presented. PFS will be plotted for each treatment group using the Kaplan-Meier method. The plot will include the number of subjects at risk for each treatment group over time.

7.2.2 Analyses of Secondary Efficacy Endpoints

Of the secondary endpoints, ORR and OS are designated as key secondary efficacy endpoints. These endpoints will be tested for statistical significance under an overall one-sided significance level of 0.025 if and only if the primary endpoint is significant. Details of multiplicity adjustment are provided in Section 6.3.

The analyses of secondary efficacy endpoints, Lymph Node Response Rate and Hematologic Improvement Rate, will include hypothesis testing, but no statistical significance will be claimed. The analyses of other secondary efficacy endpoint, Duration of Response (DOR), will be nominal only.

7.2.2.1 Overall Response Rate (ORR)

The primary analyses of ORR will use ORR based on IRC assessment in the ITT analysis set.

ORR will be derived from best overall response (BOR), which is defined as the best time point response that a subject achieves during the course of the study, with the response ranked according to the following order (from best to worst): CR>CRi>PR>PRwL>SD>PD (CRi applies to CLL only). In addition, the IRC may assign a timepoint response of unknown (UNK) due to missing, incomplete, or inadequate data; at baseline, the IRC may assign a timepoint response of no evidence of disease (NED) if both radiological and clinical data indicate no disease involvement; for a post-baseline timepoint, the IRC may also assign a timepoint response of not evaluable (NE) if no target lesions were identified at baseline and the radiological and clinical data at the post-baseline timepoint do not support the disease response of PD or UNK; these categories will be classified as OTHER in the summary.

The estimated ORR (percent of subjects with a BOR of CR, CRi, PR or PRwL) and a 2-sided 95% CI will be provided for each treatment arm. Number and percent of subjects with BOR in each of the response categories (CR, CRi, PR, PRwL, SD, PD, OTHER) will also be presented by treatment arm. All subjects in the analysis set will be included in the denominator in the calculation of the percentage for each response category or ORR.

ORR will be analyzed using the Cochran-Mantel-Haenszel (CMH) test (one-sided, see Appendix C) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. The odds ratio and its 95% CI will be estimated.

The following sensitivity analyses will be performed:

- ORR based on investigator assessment
- Overall Confirmed Response Rate (OCRR), with overall confirmed response (based on independent review) defined as best confirmed response (time between response and confirmation must be ≥8 weeks in duration) of CR, CRi, PR, or PRwL, according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Overall Response Rate without PRwL: with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR)
- Overall Confirmed Response Rate without PRwL
- Overall Response Rate in subset of subjects with baseline assessment other than UNK and NED. This analysis will be performed if there are at least 5% subjects with baseline assessment of UNK or NED
- Overall Response Rate in subset of subjects who are on treatment for 60 days or longer
- Analysis using the PP analysis set
- Analysis using the AT analysis set

Subgroup analyses will be performed for ORR using the subgroups specified in Section 6.5.

The odds ratio (duvelisib/ofatumumab) with its 95% CI will be calculated for each subgroup based on unstratified test. The odds ratios and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

For selected subgroups, the estimated ORR and its 2-sided 95% CI will be provided for each treatment arm; number and percent of subjects with BOR in each of the response categories (CR, CRi, PR, PRwL, SD, PD, OTHER) will also be presented by treatment arm.

7.2.2.2 Overall Survival (OS)

The primary analyses of OS will be based on the ITT analysis set.

Subjects without documentation of death at the time of the data cutoff for analysis will be censored at the date the subject was last known to be alive, or the data cutoff date, whichever is earlier. A stratified one-sided log-rank test will be used to compare OS between the 2 treatment groups. The HR along with the 95% CI will be estimated using a stratified Cox model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. A Kaplan-Meier plot for OS will be presented by treatment group. Estimates and 95% confidence intervals for the

25th percentile, median, and 75th percentile for OS will be presented by treatment group (if estimable). Probabilities of survival at selected time points may also be presented.

A majority of subjects can be expected to take subsequent anticancer therapy, especially after disease progression. To adjust for the effects of subsequent therapy, the following sensitivity analysis may be performed: Primary analyses except additional censoring at the start date of subsequent therapy

Subgroup analyses for OS will be performed using the subgroups specified in section 6.5. The HR (duvelisib /ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The HRs and their 95% CIs will be displayed for all subgroups graphically in a forest plot.

OS follow-up time, defined as time from randomization to the date of last procedure or assessment (including telephone interview), will be summarized by treatment arm and for the two treatment arms combined (total) using the reverse Kaplan-Meier method, which applies Kaplan-Meier method to the OS data with the same event/censoring time as the primary OS analyses while reversing event/censoring indicator. The 25th percentile, median, and 75th percentile of OS follow-up time will be presented.

7.2.2.3 Lymph Node Response Rate

The primary analyses of lymph node response rate will be based on IRC assessment in the ITT analysis set. The number of subjects with lymph node response, estimated lymph node response rate, and a 2-sided 95% CI will be provided for each treatment arm.

Lymph node response rate will be analyzed using the CMH test (one-sided) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

A sensitivity analysis will be performed using lymph node response rate based on investigator measurements.

7.2.2.4 Hematologic Improvement Rate

A subject with a hematologic improvement is one who consistently met the criteria of an improvement in neutrophil count, hemoglobin or platelet count for a period of at least 60 days during which the subject did not have a transfusion or exogenous cytokines. Missing data or missing scheduled assessments during a potential 60-day hematologic improvement episode will result in categorization of not having a hematologic improvement.

The number of subjects with hematologic improvement, estimated hematologic improvement rate, and a 2-sided 95% CI will be provided for each treatment arm.

Hematologic improvement rate will be compared between treatment groups using one-sided CMH test. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

7.2.2.5 Duration of Response (DOR)

DOR will be presented using the Kaplan-Meier method for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL) as determined by IRC assessment. The censoring method will be the same as that for the primary endpoint (see Appendix A). The Kaplan-Meier plot for DOR will be presented by treatment group. Estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for DOR will be presented by treatment group (if estimable). No treatment comparison will be performed.

7.2.3 Analyses of Exploratory Efficacy Endpoints

7.2.3.1 Improvement in Disease-Related Symptoms

Disease-related symptoms of fever, weight loss, and drenching night sweats will be assessed as present or absent at each timepoint. Improvement in each of the three disease-related symptoms from baseline will be analyzed separately. The summary will display by treatment arm the proportion of subjects with an improvement while on treatment. Descriptive analysis of improvement in these disease-related symptoms will also be presented over the course of the study (eg, for each scheduled assessment) by treatment arm.

The disease-related symptom of fatigue will be measured by the ECOG performance status. Summary statistics will be provided by treatment arm for the ECOG score and the change from baseline at each timepoint.

7.2.3.2 Minimal Residual Disease

MRD status for subjects with a response of CR or CRi will be determined using flow cytometry from a central lab. MRD will be classified as positive or negative. The estimated rates of subjects with negative MRD in subjects with a response of CR or CRi will be provided for each treatment arm given there are sufficient amount of data.

7.2.3.3 Time To Response (TTR)

TTR will be presented for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL). Summary statistics and a 2-sided confidence interval for median TTR will be provided by treatment arm. No treatment comparison will be performed.

7.3 Safety Analyses

All safety analyses will be performed using the AT analysis set.

7.3.1 Adverse Events

Adverse events will be coded using MedDRA Version 16.1 or higher. The Grade of the AE will be assessed according to the NCI-CTCAE Version 4.03. If an AE is not included in the NCI-CTCAE Version 4.03, the Grade of the AE will be assessed according to the protocol, Section ***8.2.1.2***.

The summary of AEs will be focused on treatment-emergent AEs. A treatment-emergent AE (TEAE) is defined as any AE that emerges or worsens in the period from the first dose of study treatment to 30 days after the last dose of study treatment. The onset date of an AE will be compared to the first dose date and the last dose date plus 30 days to determine whether the AE is treatment-emergent or not. If the onset date of an AE is completely or partially missing, refer to Section 6.2.1 for the algorithm to determine whether an AE is treatment emergent.

TEAEs will be summarized for each treatment arm by MedDRA system organ class (SOC) and preferred term (PT), or PT only. For summary tables by SOC and PT, SOC will be sorted alphabetically and PT will be sorted by decreasing frequency in the duvelisib arm within each SOC. For summary tables by PT only, PT will be sorted by decreasing frequency in the duvelisib arm.

If multiple TEAEs of the same PT occur within a subject, only the maximum grade observed for this PT will be used in summary of TEAEs by grade, the subject will be counted only once in the number of subjects for this PT and only once for the number of subjects for the SOC to which this PT belongs.

An overview TEAE summary table will be provided, which will include the number of subjects with AEs in selected categories. In addition, TEAEs will be summarized for the following categories, and will be tabulated by SOC and PT, unless otherwise specified.

- Treatment-emergent AEs (All Causalities)
- Treatment-emergent AEs (Treatment-Related)
- Treatment-emergent AEs (All Causalities, by maximum grade)
- Treatment-emergent AEs (Treatment-Related, by maximum grade)
- Grade 3 or higher treatment-emergent AEs (All Causalities)
- Grade 3 or higher treatment-emergent AEs (Treatment-Related)
- Treatment-emergent SAE (All Causalities)
- Treatment-emergent SAE (Treatment-Related)
- Treatment-emergent AEs resulting in discontinuation of study drug (All Causalities)
- Treatment-emergent AEs resulting in dose hold (All Causalities)
- Treatment-emergent AEs resulting in dose reduction (All Causalities)
- Treatment-emergent AEs resulting in dose hold or reduction (All Causalities)
- Treatment-emergent AEs resulting in discontinuation of study drug (Treatment-Related)
- Treatment-emergent AEs resulting in death (All Causalities)
- Treatment-emergent AEs resulting in death (Treatment-Related)
- Treatment-emergent AEs by PT (All Causalities)

- Treatment-emergent AEs by PT (AEs with onset date within 24 weeks after the first dose of study treatment; All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (AEs with onset date within 24 weeks after the first dose of study treatment; All Causalities)
- Treatment-emergent SAEs by PT (All Causalities)
- Treatment-emergent SAEs by PT (Treatment-Related)
- Treatment-emergent AEs reported in ≥5% of subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
- Grade 3 or higher treatment-emergent AEs reported in ≥5% of subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)

A by-subject listing of the following AE categories will be presented.

- All AEs (TEAEs will be flagged)
- All SAEs (TEAEs will be flagged)
- Treatment-emergent AEs resulting in dose hold
- Treatment-emergent AEs resulting in dose reduction
- Treatment-emergent AEs resulting in discontinuation of study drug
- Treatment-emergent AEs resulting in death

Treatment-emergent AEs (All Causalities) and Treatment-emergent AEs (Treatment-Related) will also be tabulated in each subgroup specified in Section 6.5.

7.3.2 Laboratory Data

Laboratory tests will be reported separately for hematology and blood chemistry.

For the purposes of presentation in both tables and listings, the following laboratory test results will be converted to the International System of Units (SI) before presentation: sodium, potassium, chloride, bicarbonate (or CO_2), albumin, total protein, creatinine, uric acid, calcium, phosphorus, magnesium, glucose, total and direct bilirubin, and alkaline phosphatase, red blood cell (RBC) count, hemoglobin, hematocrit, platelets, white blood cell count with 5-part differential performed manually or by flow cytometry (lymphocytes, neutrophils, monocytes, basophils, and eosinophils), etc.

Lab tests performed after the start of the first dose of study treatment, up to 30 days from the last dose, will be considered treatment-emergent. Only treatment-emergent lab tests will be included in the analyses in this section.

If a laboratory test value is reported using a non-numeric qualifier (e.g., less than [<] a certain value, or greater than [>] a certain value), the given numeric value will be used in the summary statistics, ignoring the non-numeric qualifier.

For laboratory tests with NCI-CTCAE grades, a shift table from baseline grade to the maximum post-baseline grade will be provided. Laboratory tests with bi-directional grades (e.g., Hyperglycemia and Hypoglycemia) will be presented separately for each direction within the shift table.

Listings will be provided for all laboratory test results and for laboratory test results grade 3 and higher. A listing of subjects with ALT or AST >3xULN with simultaneous total bilirubin >2xULN will be presented, where ULN stands for upper limit of normal. The elevations of ALT/AST and total bilirubin must occur within 2 days of each other.

7.3.3 Vital Signs

The actual values of vital sign parameters, including temperature, heart rate, weight and systolic and diastolic blood pressure, will be presented in a by-subject listing.

7.3.4 Electrocardiogram (ECG)

A by-subject listing will be presented for baseline ECGs and post-baseline unscheduled ECGs (if any).

7.3.5 Concomitant Medications and Procedures

Please refer to Section 7.1.7 for the definition and summary of concomitant medications.

Concomitant procedures will not be summarized. A by-subject listing will be presented.

7.3.6 Deaths

On-treatment deaths are defined as deaths that occur within 30 days of the last dose of study medication. Deaths on treatment and causes for deaths occurring on-treatment, and deaths during follow-up will be summarized.

A by-subject listing of deaths on study will be presented.

7.4 Quality of Life Instruments

Analyses of the QoL instruments (EQ-5D and FACIT-F) will be performed on the ITT analysis set. Additional analyses may be considered as needed.

7.4.1 EQ-5D

The EQ-5D contains a descriptive system with one response for each of 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). The first response is coded as 1 (indicating no problems), the second response is coded as a 2 (indicating some problems), and the third response is coded as a 3 (indicating extreme problems). Ambiguous responses (eg, more than one response in a dimension) are treated as missing values. The EQ-5D

also contains a Visual Analog Scale (VAS) for health state ranging from 0 to 100, where 0 represents worst imaginable health state and 100 represents the best imaginable health state.

For each dimension of the descriptive system, the number and percentage of subjects with no problems, some problems, and extreme problems will be reported at each visit.

Summary statistics for the VAS of health state and the change from baseline will be reported for each visit.

7.4.2 FACIT-F

The FACIT-F contains 5 subscales, Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, Functional Well-Being, and Fatigue (Additional Concerns on eCRF). Each subscale contains multiple items, and each item has a response of 0 to 4. Scores will be computed for each subscale, as well as for the FACIT-F Trial Outcome Index, FACT-G total score, and the FACIT-F total score.

The Physical Well-Being subscale contains 7 items. To compute the subscale score, compute the item score by subtracting each item response from 4 (item score = 4 – item response), sum the item scores, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Physical Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Social/Family Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Social/Family Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Emotional Well-Being subscale contains 6 items. To compute the subscale score, compute the item score first. For items 1, 3, 4, 5, and 6, subtract each item response from 4 (item score = 4 – item response). For item 2, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 6 and divide by the number of items answered to compute the subscale score. At least 3 questions must be answered. If fewer than 3 questions are answered, then the subscale score will be missing. The Emotional Well-Being score ranges from 0 to 24 with higher scores representing better quality of life.

The Functional Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Functional Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Fatigue subscale contains 13 items. To compute the subscale score, compute the item score first. For all items except An5 and An7, subtract each item response from 4 (item score = 4 – item response). For items An5 and An7, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 13 and divide by the number of

items answered to compute the subscale score. At least 7 questions must be answered. If fewer than 7 questions are answered, then the subscale score will be missing. The Fatigue score ranges from 0 to 52 with higher scores representing better quality of life.

To derive the FACIT-F Trial Outcome Index, sum the subscale scores for Physical Well-Being, Functional Well-Being, and Fatigue. If any of these subscale scores is missing, the score for the FACIT-F Trial Outcome Index will be missing. The score ranges from 0 to 108.

To derive the FACT-G score, sum the subscale scores for Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, and Functional Well-Being. If any of these subscale scores is missing, the FACT-G score will be missing. The score ranges from 0 to 108.

To derive the FACIT-F total score, sum all 5 subscale scores. If any subscale score is missing, the FACIT-F total score will be missing. The score ranges from 0 to 160.

For each of the five subscale scores and for each of the FACIT-F Trial Outcome Index, the FACT-G score, and the FACIT-F total score, summary statistics for the actual values and changes from Baseline will be reported at each visit.

7.5 Pharmacokinetic Analyses

Plasma samples will be analyzed for duvelisib and IPI-656 (metabolite) concentrations using a validated high performance liquid chromatography-tandem mass spectrometry (LC/MS/MS) method. A listing of the duvelisib and IPI-656 concentrations will be provided; the sample collection date and time, and the date, time and amount of the preceding duvelisib dose administration will be included in the listing.

The PK data collected will be analyzed by standard population PK methods, using appropriate software. Analysis of exposure-response relationships for efficacy and safety endpoints will be conducted. If there is only a limited amount of plasma concentration data from this study, the data may be pooled with the results of other studies to perform the population PK and exposure-response analyses. Further details on these analyses will be outlined in a separate analysis plan. Results of the population PK and exposure-response analysis will be summarized in a separate technical report.

7.6 Biomarker Analyses

The following exploratory analyses may or may not be presented formally. If presented, they will be limited to biomarkers for which data from a sufficient number of subjects are available. Further details on these analyses may be outlined in a separate analysis plan, and results may be summarized in a separate technical report.

7.6.1 Pharmacodynamic and Potentially Predictive Biomarkers

The relationship between tumor genomics, gene expression, protein expression, and clinical endpoints may be explored as follows:

- Evaluation of serum biomarkers, blood immunophenotype and tumor biomarkers for relationships with duvelisib clinical activity and safety
- Evaluation of tumor genomic features (eg, cytogenetics, FISH, DNA sequence variation, DNA copy number variation, and/or RNA expression) for relationships with duvelisib clinical activity
- Evaluation of the relationship between serum and/or tumor biomarkers and disease progression of duvelisib and ofatumumab

7.6.2 Pharmacogenomics

DNA may be extracted from the pharmacogenomics sample in order to evaluate a subject's germline DNA. This sample may be used to explore the following:

- Evaluate the relationship between germline DNA sequence variations and PK of duvelisib
- Evaluate germline DNA as a control to verify that tumor sequence variations are somatic mutations

8 CHANGES IN PLANNED ANALYSES

8.1 Changes in Planned Analyses from Protocol Amendment 1 to SAP V1.0

There are no changes from Protocol Amendment 1. More details are provided in the SAP.

8.2 Changes in Planned Analyses from SAP V1.0 to SAP V2.0

The main changes from SAP V1.0 to SAP V2.0 are described below.

1. Section 3.5.2 Clarification

   Old (V1.0): The second analysis of OS will be performed at the PFS final analysis, which is expected to happen after 185 PFS events have occurred, or approximately 27.5 months after the first subject is randomized.

   New (V2.0): The second interim analysis of OS will be performed at the PFS final analysis, which is to happen after 185 PFS events have occurred. If the planned final analysis of PFS is not performed due to early stop for efficacy, then the second interim analysis of OS will be conducted approximately one year after the first interim analysis of OS.

2. New censoring method for event-free survival were specified.
3. Prior irradiation and prior surgery were removed from prior therapy summary.

The following additional analyses/summaries have been added to this version of the SAP:

1. Analysis of ORR of subjects who are on treatment for 60 days or longer
2. Time to OS follow up
3. Time to response
4. Treatment-emergent AEs reported in ≥5% of CLL/SLL subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
5. Grade 3 or higher treatment-emergent AEs reported in ≥5% of CLL/SLL subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
6. Treatment-emergent AEs by PT (All Causalities, AE onset date within 24 weeks after the first dose of study treatment)
7. Grade 3 or higher treatment-emergent AEs by PT (All Causalities, AEs with onset date within 24 weeks after the first dose of study treatment)
8. Treatment-emergent AEs resulting in discontinuation of study drug (Treatment-Related)
9. Treatment-emergent AEs resulting in death (Treatment-Related)
10. Treatment-emergent SAEs by PT (All Causalities).
11. Treatment-emergent SAEs by PT (Treatment-Related)
12. Deaths on treatment
13. Deaths during follow-up

14. A by-subject listing of treatment-emergent AEs resulting in dose hold
15. A by-subject listing of treatment-emergent AEs resulting in dose reduction
16. A by-subject listing of deaths on study
17. A summary table and a listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF
18. Comparision of del[17p] values captured from central lab and local lab
19. The following biomarkers will be summarized as baseline characteristics:
    - del[17p] (captured from central lab, presence vs absence vs indeterminate)
    - del[17p] (captured from local lab, presence vs absence vs indeterminate)
    - del[17p] and/or TP53 mutation (del[17p] and/or TP53 mutation, del[17p] captured from central lab, either or both present vs neither present vs indeterminate)
    - del[17p] and TP53 mutation (del[17p] and TP53 mutation, del[17p] captured from central lab, both present vs either vs neither present vs indeterminate)
    - TP53 mutation (presence vs absence vs indeterminate)
    - IgHV status (mutated vs unmutated vs indeterminate)
    - CD38 (Positive (≥30%) vs Negative vs indeterminate)
    - ZAP70 (Positive (>19%) vs Negative (<=19%) vs indeterminate)
20. Subgroup analyses of PFS, OS and ORR based on the following factors:
    - Time from last dose of most recent prior anti-cancer therapy to randomization <12 months: yes vs no
    - del[17p] (captured from central lab) or TP53 mutation (either or both present vs neither present)

8.3 Changes in Planned Analyses from SAP V2.0 to SAP V3.0

The main changes from SAP V2.0 to SAP V3.0 are described below.

1. In Section 3.5.2, the estimated numbers of OS events at each of OS interim analyses and final analysis are provided.
2. In Section 6.3, Lymph Node Response Rate and Hematological Improvement Rate are removed from the list of key secondary efficacy endpoints and removed from the hierarchy of inferential testing in secondary endpoints.
3. In Section 7.2.1.3, for selected subgroups, additional analyses of PFS are added. These include the estimate and 95% CI for the median by treatment group, and the Kaplan-Meier plot by treatment group.
4. In Section 7.2.2.1, for selected subgroups, summaries of BOR, estimated ORR and its 2-sided 95% CI for each treatment group are added.

9 REFERENCES

1 O'Brien PC, Fleming TR. *A multiple testing procedure for clinical trials.* Biometrics 1979; 35:549-556.

2 Robins JM. *Information recovery and bias adjustment in proportional hazard regression analysis of randomized trials using surrogate markers*. Proceedings of the Biopharmaceutical Section, American Statistical Association 1993; pp. 24-33.

3 Robins JM, Finkelstein D. *Correcting for Non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests*. Biometrics, 2000; 56(3):779-788.

4 Robins JM, Tsiatis, AA. *Correcting for noncompliance in randomized trials using rank preserving structural failure time models.* Communications in Statistics, 1991; 20, 2609-2631.

10 APPENDICES

10.1 **Appendix A: Primary PFS Event/Censoring Method**

Censoring of PFS will be performed as detailed in the table below.

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression or death before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death	Date of death	Event
Documented progression or death following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment or procedure started before documented progression	Date of last adequate disease status assessment prior to new anticancer treatment	Censored

Note: Disease status assessment includes CT scans (chest, abdomen and pelvis), bone marrow aspirate and/or biopsy (may not be required of all subjects at all scheduled disease status assessments), CBC and differential count, focused physical examination, disease related constitutional symptoms for disease assessment, and ECOG performance status. An adequate baseline disease status assessment is any baseline disease status assessment that include WBC counts and CT scans. An adequate post-baseline disease status assessment is any disease status assessment for which a disease status (eg, CR, CRi, PR, PRwL, SD, and PD) is arrived per protocol-defined criteria by the IRC (for IRC assessment) or investigator (for investigator assessment).

10.2 Appendix B: **Event-Free Survival (EFS) Event/Censoring Method**

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death or new anticancer treatment/procedure occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression, death or new anticancer treatment/procedure started before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression, and no new anticancer treatment/ procedure started before documented progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death, and no new anticancer treatment/procedure started before death	Date of death	Event
Documented progression, death or new anticancer treatment/procedure following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment/procedure started before documented progression with ≤1 missing scheduled disease status assessment prior to new anticancer treatment/procedure	Start date of new anticancer treatment/procedure	Event

10.3 **Appendix C: One-Sided Cochran–Mantel–Haenszel (CMH) Test**

The Cochran–Mantel–Haenszel (CMH) test compares binary responses of two treatment groups, adjusting for stratification factors. In the CMH test, the data are arranged in a series of associated 2 × 2 contingency tables, the null hypothesis is that the observed response is independent of the treatment used in any 2 × 2 contingency table.

Let O_{hij} be the observed frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h and E_{hij} be the expected frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h, where

$$E_{hij} = \frac{(O_{hi1} + O_{hi2}) \cdot (O_{h1j} + O_{h2j})}{O_{h11} + O_{h12} + O_{h21} + O_{h22}}$$

Also, let P_{hij} be the probability of outcome being j given treatment being i (i=1 or 2, j= 1 or 2) in stratum h. To test the following hypothesis,

$$H_0 : P_{h11} = P_{h21} \text{ for all } h \in \{1, \ldots, H\}$$

versus

$$H_1 : P_{h11} > P_{h21} \text{ for at least one } h \in \{1, \ldots, H\}$$

The one-sided Cochran-Mantel-Haenszel Test Statistics is constructed as:

$$Z_{CMH} = \frac{\sum_{h=1}^{H} (O_{h11} - E_{h11})}{\sqrt{V_{11}}} \sim Z$$

Where

$$V_{11} = Var(\sum_{h=1}^{H} (O_{h11} - E_{h11})) = \sum_{h=1}^{H} \frac{E_{h11} \bullet E_{h22}}{O_{h11} + O_{h12} + O_{h21} + O_{h22} - 1}$$

The test statistic Z_{CMH} will be compared with standard normal distribution to obtain p-value of the CMH test.

STATISTICAL ANALYSIS PLAN

Protocol IPI-145-07

A Phase 3 Study of IPI-145 versus Ofatumumab in Subjects with Relapsed or Refractory Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma

Protocol Version:	**Amendment 2**
Type of Analysis Plan:	**Interim and Final Analyses**
Version:	**3.0**
Date:	**13 Apr 2016**
Author:	**Shijie Tang, PhD**

STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

The undersigned has developed this statistical analysis plan (SAP):

Name/Title	Signature	Date
Shijie Tang, PhD Associate Director Biostatistics		14 Apr 2016

The undersigned have reviewed this SAP and approve it in its entirety:

Name/Title	Signature	Date
Virginia Kelly, MD Senior Medical Director Clinical Development	(Claudia Dansky Ullman for V. Kelly)	14 April 2016
Rekha Garg, MD, MS Vice President Safety and Risk Management	Rekha Garg	14 Apr 2016
Justin McLaughlin, MA Director Medical Writing		14 April 2016
Lixin Han, PhD Senior Director Biostatistics		13 Apr 2016

DOCUMENT HISTORY

Version	Date	Author(s)	Brief Summary of Changes
1.0	02 Oct 2014	Shijie Tang	Original
2.0	24 Nov 2015	Shijie Tang	Clarified the timing of the 2nd interim analysis of overall survival; added biomarkers to baseline characteristics; other changes. Refer to Section 8.2 for Details.
3.0	06 Apr 2016	Shijie Tang	In Section 3.5.2, the estimated numbers of OS events at each of OS interim analyses and final analysis are provided. Lymph node response rate and hematological improvement rate are removed from the hierarchy of inferential testing in secondary endpoints.

Table of Contents

LIST OF ABBREVIATIONS

Abbreviation	Description
ADaM	Analysis Data Model
AE	Adverse Event
ANCOVA	Analysis of Covariance
AT	All-Treated
ALT	Alanine Aminotransferase
ANC	Absolute Neutrophil Count
AST	Aspartate Aminotransferase
ATC	Anatomical Therapeutic Chemical
BID	Twice a day
BOR	Best Overall Response
BUN	Blood Urea Nitrogen
CI	Confidence Interval
CLL	Chronic Lymphocytic Leukemia
CMH	Cochran–Mantel–Haenszel
CO2	Bicarbonate
CR	Complete Response
CRi	Complete Response with Incomplete Marrow Recovery
CS	Abnormal and Clinically Significant
CTCAE	Common Terminology Criteria for Adverse Events
DOR	Duration of Response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form
eDISH	Evaluation of Drug Induced Serious Hepatotoxicity
EQ-5D	European Quality of Life-5 Dimensions
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue
FACT-G	The Functional Assessment of Cancer Therapy - General

Abbreviation	Description
HR	Hazard Ratio
IDMC	Independent Data Monitoring Committee
IPI-145	(S)-3-(1-(9H-purin-6-ylamino)ethyl)-8-chloro-2-phenylisoquinolin-1(2H)-one
IPCW	Inverse Probability of Censoring Weighted
IRC	Independent Review Committee
IRT	Interactive Response Technology
ITT	Intent-To-Treat
IV	Intravenous
IWCLL	International Workshop on Chronic Lymphocytic Leukemia
IWG	International Working Group
LC/MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LDH	Lactate Dehydrogenase
LFT	Liver Function Test
MedDRA	Medical Dictionary for Regulatory Activities
MEOI	Medical Events of interest
MMRM	Mixed Effect Repeated Measures
MRD	Minimal Residual Disease
NCI	National Cancer Institute
NCS	Abnormal but Not Clinically Significant
NED	No Evidence of Disease
ORR	Overall Response Rate
OS	Overall Survival
PD	Progressive Disease
PFS	Progression-Free Survival
PK	Pharmacokinetics
PP	Per-Protocol
PR	Partial Response
PRwL	PR with Lymphocytosis

Abbreviation	Description
PT	Preferred Term
QoL	Quality of Life
RBC	Red Blood Cell
RPSFT	Rank Preserving Structural Failure Time
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
SI	Standard International System of Units
SLL	Small Lymphocytic Lymphoma
SOC	System Organ Class
SPD	Sum of Products
TEAE	Treatment Emergent Adverse Event
TTR	Time to Response
UNK	Unknown
VAS	Visual Analog Scale
WHO	World Health Organization
WHO-DD	World Health Organization Drug Dictionary

1 INTRODUCTION

This is the statistical analysis plan (SAP) for study IPI-145-07, *A phase 3 study of IPI-145 versus ofatumumab in subjects with relapsed or refractory chronic lymphocytic leukemia/small lymphocytic lymphoma*. This SAP is prepared according to Amendment 2 of the protocol, dated 02 March 2015.

2 STUDY OBJECTIVES

2.1 Primary Objective

The primary objective of this trial is to examine the efficacy of duvelisib (IPI-145) monotherapy versus ofatumumab monotherapy in subjects with relapsed or refractory Chronic Lymphocytic Leukemia (CLL) or Small Lymphocytic Lymphoma (SLL).

2.2 Secondary Objectives

- To determine the safety of duvelisib in subjects with CLL or SLL
- To evaluate the pharmacokinetics (PK) of duvelisib and, if applicable, its metabolite(s)

2.3 Exploratory Objectives

- To evaluate the health-related quality of life (QoL) of subjects
- To evaluate pharmacodynamic biomarkers of duvelisib
- To evaluate biomarkers that may predict duvelisib clinical activity and/or safety
- To evaluate mechanisms of resistance in subjects who exhibit disease progression while being treated with duvelisib or ofatumumab
- To evaluate genomic features of tumors predictive of response in subjects treated with duvelisib or ofatumumab

3 STUDY DESIGN

3.1 Overview

Study IPI-145-07 is a two-arm, randomized (1:1 ratio), open-label, phase 3, superiority trial designed to evaluate the efficacy and safety of duvelisib compared to ofatumumab.

Subjects who meet all the eligibility criteria at Screening will return to the clinic on Day 1 to receive their first dose of study drug (randomized to either duvelisib or ofatumumab). The first treatment cycle for each treatment arm will be 3 weeks (21±2 days). Subsequent treatment cycles will be 4 weeks (28±4 days).

Subjects randomized to duvelisib will be given a starting dose of 25 mg duvelisib administered orally twice daily (BID) initially in a 21-day treatment cycle followed by 28-day treatment cycles for up to 18 cycles or until disease progression or unacceptable toxicity (whichever comes first). After completing approximately 18 cycles of treatment, subjects may receive additional cycles of duvelisib for up to 3 years (39 cycles of total treatment) if they have documented evidence of response and disease requiring continued treatment according to the modified IWCLL/revised IWG criteria.

Subjects randomized to ofatumumab will receive treatment consistent with the approved product labeling: 8 weekly infusions, starting with an initial IV dose of 300 mg ofatumumab on Day 1 followed by seven weekly doses of 2000 mg. Thereafter, subjects will receive 2000 mg ofatumumab once every month for four months or until disease progression or unacceptable toxicity (whichever comes first), as outlined in the schedule of assessments. Administration of ofatumumab will not exceed the 12 doses (within 7 cycles) as described in the prescribing information.

Subjects will be followed for survival for up to 6 years (or other duration as specified by the protocol) from randomization or until death. Follow-up visits will occur every 6 months which can be conducted through a telephone interview.

3.2 Sample Size Consideration

This study employs a randomized, open-label, parallel design to assess the potential superiority of duvelisib treatment over ofatumumab treatment on progression-free survival (PFS) in CLL or SLL subjects.

The sample size was determined to provide sufficient power to test the primary endpoint of PFS in a group sequential design with one planned interim analysis.

Assuming an exponential distribution for PFS, a total of 185 PFS events (deaths or progressions) will provide approximately 93% power to detect a hazard ratio of 0.6 (median PFS of 9 months in the ofatumumab group versus 15 months in the duvelisib group) using a one-sided log-rank test at a 2.5% overall significance level (ie, alpha of 0.025). The study design employs the Lan-DeMets spending function for O'Brien-Fleming boundary as the alpha spending function and the Hwang-Shih-DeCani gamma (-4) spending function as the beta spending function. A total of 300

subjects will be randomized in a 1:1 ratio to receive either ofatumumab or duvelisib. Assuming a peak accrual rate of 30 subjects per month with 11 months for accrual ramp-up and a 4% cumulative dropout rate per year, the enrollment would complete in 16 months, with the final analysis for PFS estimated to occur approximately 27.5 months from the randomization of the first subject.

3.3 Randomization

Once a subject has met all entry criteria, the Interactive Response Technology (IRT) will be used to generate a distinct subject identifier. If a subject discontinues from the study, the subject identifier will not be reused and the subject will not be allowed to re-enter the study.

Eligible subjects will be randomized via IRT in a 1:1 ratio to one of two treatment arms:

- Arm 1: duvelisib 25 mg BID
- Arm 2: Ofatumumab

In order to ensure subject balance between treatment groups, study subjects will be stratified by the following:

- High risk cytogenetics (presence vs absence of del[17p])
- Refractory/early relapse to purine analog based treatment (progression <12 months after fludarabine/pentostatin: yes vs no)
- Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)

First dose is to occur within 7 days of randomization after all screening assessments have been completed.

3.4 Blinding

This study is open-label and Investigators, site staff, Sponsor and Sponsor designees will have access to treatment assignments for individual subjects. To reduce bias, a blinded independent central review of disease status will be conducted, which will be blinded to individual subject treatment assignments. Investigators, site staff, Sponsor and Sponsor designees will only have access to blinded (pooled) aggregate study data, except for a select number of Sponsor staff, who may review aggregate SAEs and Medical Events of Interest (MEOI) data by treatment groups.

3.5 Planned Analyses

Interim and final analyses are planned for PFS and OS. For an interim analysis, the actual p-value boundaries will be calculated based on the actual number of PFS and OS events at the analysis. For the final analysis, the actual p-value boundary for efficacy will be calculated based on the actual number of PFS and OS events at the interim analysis (or analyses) and the final analysis.

The Independent Data Monitoring Committee (IDMC) will review PFS, OS and other efficacy data at the time of PFS interim analysis. The list of IDMC deliverables, including tables and listings, is included in the IDMC charter. These deliverables will be based on the methods described in this SAP.

3.5.1 Planned Analyses for PFS

There will be two planned analyses for PFS, one interim analysis and one final analysis. After approximately 50% of the planned PFS events (ie, 93 PFS events) have been observed based on the blinded, independent, central review, an interim analysis of efficacy will be performed.

At the interim analysis, PFS will be tested at the alpha level based on the Lan-DeMets alpha spending function for the O'Brien-Fleming boundary with the opportunity to stop the study (other than survival follow-up) for overwhelming evidence of efficacy. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for efficacy is one-sided p-value of 0.0015 (corresponding approximately to a HR of 0.540). In the meantime, PFS will also be tested at the alpha level for futility based on the alpha and beta spending functions specified for the study design. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for futility is one-sided p-value of 0.4791 (corresponding approximately to a HR of 0.990). The futility boundary of this study is non-binding, meaning that the Type I error is properly controlled even if the study is continued after the futility boundary for PFS is crossed at the interim analysis. Under the protocol design assumptions (eg, the study is fully enrolled in 16 months), the interim analysis is expected to occur approximately 17 months after the first subject is randomized.

If the study is not stopped at the interim analysis, the final analysis will be performed when approximately 185 PFS events have occurred in the study. If the interim analysis and the final analysis occur at exactly 93 and 185 PFS events, the criterion of statistical significance (ie, boundary for efficacy) is one-sided p-value of 0.0245 (corresponding approximately to a HR of 0.748).

3.5.2 Planned Analyses for OS

There will be three planned analyses for OS: two interim analyses and one final analysis. The first interim analysis of OS will be performed at the time of the planned PFS interim analysis after 93 PFS events have occurred. The estimated number of OS events at this first OS interim analysis is 24. The second interim analysis of OS will be performed at the planned PFS final analysis, after 185 PFS events have occurred. If the planned final analysis of PFS is not performed due to an early stop for efficacy, then the second interim analysis of OS will be conducted approximately 1 year after the first interim OS analysis. The estimated number of OS events at the second OS interim analysis is 58. The final analysis of OS will take place after the completion of follow-up for all subjects. The estimated number of OS events at the final OS analysis is 161. The information fractions at the two interim OS analyses and the final OS analysis will be calculated based on 161 planned OS events at the final OS analysis. The cumulative alpha spending at the three analyses will be 0.0001, 0.0249, and 0.0250, respectively.

If the OS analyses occur at 24, 58, and 161 OS events, the p-value stopping boundaries for the three analyses will be 0.0001, 0.0249 and 0.0003, respectively. The actual p-value stopping boundaries will be derived from the actual number of OS events that occurred. The statistical significance of OS will be claimed based on the hierarchical testing procedure as specified in Section 6.3.

4 ANALYSIS SETS

4.1 Intent-To-Treat (ITT) Analysis Set

The intent-to-treat (ITT) analysis set includes all subjects who are randomized, with treatment group designated according to randomization.

This analysis set will serve as the primary analysis set for all efficacy endpoints and demographics.

4.2 All-Treated (AT) Analysis Set

The all-treated (AT) analysis set includes all subjects who receive any amount of study drug (duvelisib or ofatumumab), with treatment group designated according to actual study treatment received. The AT analysis set will be the primary analysis set for safety endpoints.

4.3 Per-Protocol (PP) Analysis Set

The per-protocol (PP) analysis set includes all subjects in the ITT analysis set who do not violate the terms of the protocol in a way that would significantly affect the study outcome, with treatment group designated according to randomization. Subjects who meet any of the following criteria may be excluded from this analysis set:

- Do not have documented diagnosis of CLL or SLL
- Do not have measurable nodal disease at baseline as determined by the IRC
- ECOG performance status >2
- History of Richter's transformation or prolymphocytic leukemia
- Refractory to ofatumumab (defined as progression or relapse <12 months of receiving ofatumumab monotherapy or <24 months of receiving an ofatumumab-containing regimen)
- Prior exposure to a PI3K inhibitor
- Receive concomitant prohibited anticancer therapy
- Permanent discontinuation from study drug due to non-compliance

The PP analysis set will be a secondary analysis set for selected efficacy analyses.

5 STUDY ENDPOINTS

5.1 Primary Endpoint

The primary efficacy endpoint PFS is defined as time from randomization to the first documentation of progressive disease (PD) as determined by independent review or death due to any cause.

The censoring method of the primary endpoint can be found in Appendix A. The detailed algorithm will be specified in the Analysis Data Model (ADaM) specifications.

5.2 Secondary Endpoints

The secondary efficacy endpoints of the study are:

- Overall Response Rate (ORR), with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR), or PR with lymphocytosis (PRwL), according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Overall Survival (OS), defined as time from randomization to death
- Lymph node response rate, with lymph node response defined as ≥50% decrease in the SPD of target lymph nodes
- Hematologic improvement rate, defined as any of following hematologic improvement sustained for at least 60 days without transfusion or exogenous growth factors:
    - Neutrophil count >1,500/µL or an increase ≥50% from Baseline; or
    - Hemoglobin >11 g/dL or an increase ≥50% from Baseline; or
    - Platelet count >100,000/µL or an increase ≥50% from Baseline
- Duration of Response (DOR), defined as time from the first documentation of response to first documentation of PD or death due to any cause.

Other secondary endpoints of the study are:

- TEAEs and changes in safety laboratory values
- PK parameters derived from plasma duvelisib concentrations and, if applicable, its metabolite(s)

5.3 Exploratory Endpoints

The exploratory efficacy endpoints of the study are:

- Improvement in disease-related symptoms of fatigue, weight loss, unexplained fevers and night sweats from Baseline
- Minimal Residual Disease (MRD) in subjects with documented CR or CRi
- Time to Response (TTR), defined as the time from randomization to the first documentation of response (CR, CRi, PR or PRwL)

The QoL endpoints of the study are:

- Health-related QoL of subjects as assessed by the following subject-reported questionnaires:
    - EuroQol-5D (EQ-5D)
    - Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F)

The biomarkers of the study are:

- Serum and tissue biomarkers and blood immunophenotype
- Tumor genomic features (eg, DNA sequence variation, DNA copy number variation, and/or RNA expression)
- Germline DNA sequence variations

6 GENERAL STATISTICAL METHODS AND DATA HANDLING

6.1 General Methods

Summary statistics will be presented by treatment group, unless stated otherwise.

Unless otherwise specified, descriptive statistics for continuous data will include the number of subjects with data to be summarized (n), mean, standard deviation, 25% quartile, median, 75% quartile, and minimum and maximum. The same number of decimal places as in the raw data will be presented when reporting the minimum and maximum, one more decimal place than the raw data will be presented when reporting mean and median, and 2 more decimal places than the raw data will be presented when reporting standard deviation.

Descriptive statistics for categorical/qualitative data will include frequency counts and percentages. The total number of subjects in the treatment group will be used as the denominator for percent calculations, unless stated otherwise. All percentages will be presented with one decimal, unless otherwise specified. Percentages equal to 100 will be presented as 100, and percentages will not be presented for zero frequencies.

Descriptive statistics associated with time-to-event analyses will include the number of events, the number of subjects censored, 25% quartile, median, 75% quartile, and 95% confidence interval for median. These statistics will be presented for all time-to-event analyses, unless stated otherwise.

Listings will be provided for selected endpoints. For listings broken down by center and treatment arm, site (center) number will be ordered by country.

6.2 Handling of Missing Data

In general, values for missing data will not be imputed unless methods for handling missing data are specified.

6.2.1 Handling of Missing Dates/Months/Years for Adverse Events

Adverse events (AEs) with incomplete onset dates will be handled as follows for the sole purpose of determining treatment emergence (TEAE is defined in Section 7.3.1):

- If the start/end date of an AE is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The AE will be assumed to be treatment emergent if it cannot be definitively shown that the AE did not occur or worsen during the treatment-emergent period (worst case approach). The detailed algorithm will be specified in ADaM specifications.
- If the start date is completely missing, an AE will be considered treatment-emergent unless the stop date is before study drug administration.
- If the dose start date is missing for a subject at a data-cut, all AEs of the subject will be considered treatment-emergent.

The original partial or missing date will be shown in listings of AEs.

6.2.2 Handling of Missing Dates/Months/Years for Concomitant Medications

Prior or concomitant medications with incomplete start dates will be handled as follows for the sole purpose of determining whether a non-study medication is a concomitant medication:

- If the start/stop date of a medication is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The medication will be assumed to be concomitant if it cannot be definitively shown that the stop date is before the start of administration of study drug, or the start date is more than 30 days after the last date of administration of study drug. The detailed algorithm will be specified in ADaM specifications.
- If the start/stop dates are completely missing, a medication will be considered concomitant.
- If the dose start date is missing for a subject at a data-cut, all non-study medications of the subject will be considered concomitant.

The original partial or missing date will be shown in listings of all non-study medications.

6.2.3 Handling of Missing Dates/Months/Years for Disease History and Prior Therapies

For the purpose of calculating the duration from initial diagnosis, most recent relapse/refractory diagnosis or most recent prior therapy to randomization, partial/missing dates for diagnosis and last prior therapy completion will be imputed as follows:

- If both date and month are missing and the year is prior to the year of screening, the imputed date and month will be 01 July.
- If both date and month are missing and the year is the same as the year of screening, the imputed date will be the middle point between 01 Jan of the year and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- If date is missing and the month and year are prior to the month and year of screening, the imputed date will be 15th day of the month.
- If date is missing and the month and year are the same as the month and year of screening, the imputed date will be the middle point between the first date of the month and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- No imputation will be performed if the year is missing.

6.3 Multiple Comparisons/Multiplicity Adjustment

Of the secondary endpoints, ORR and OS are designated as key secondary efficacy endpoints. The primary endpoint and key secondary endpoints will be tested at an overall one-sided alpha level of 0.025 based on a gatekeeping approach.

If the primary endpoint is significant, the two key secondary endpoints will be sequentially tested in the order listed above. ORR will be tested at the one-sided 0.025 level only if PFS is declared statistically significant. OS will be tested according to the planned alpha spending specified in

Section 3.5.2 only if PFS and ORR are declared statistically significant. If any null hypothesis is not rejected in this sequence of tests, formal sequential testing will be stopped.

6.4 Adjustments for Covariates

Adjustments for covariates will be considered for analysis of primary and key secondary endpoints, with details provided in Sections 7.2.1 and 7.2.2.

6.5 Subgroups

PFS, ORR and OS will be examined in the subgroups based on the following baseline characteristics or variables:

- Stratification factors (1. captured from central lab; 2 and 3 captured on the eCRF):
    1 High-risk cytogenetics (presence vs absence of del[17p])
    2 Refractory/early relapse to purine analog-based therapy (defined as progression <12 months after fludarabine/pentostatin: yes vs no)
    3 Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)
- Diagnosis (CLL or SLL)
- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Previously treated with ofatumumab (yes vs no, if sample size allows the analysis)
- Time from last dose of most recent prior anti-cancer therapy to randomization <12 months: yes vs no
- del[17p] or TP53 mutation (either or both present vs neither present; del[17p] will be as captured from central lab)

TEAEs (All Causalities) and TEAEs (Treatment-Related) will be examined in the following subgroups:

- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Baseline ECOG performance status (0 or 1 vs 2)

More details will be specified in Sections 7.2.1, 7.2.2 and 7.3.1.

6.6 Visit Windows

All data will be categorized based on the scheduled visit at which it was collected. These visit designators are predefined values that appear as part of the visit tab in the eCRF. There will be no additional analysis windowing done based on the assessment date.

6.7 Unscheduled Visits

Unscheduled visits will not be included in by-visit summary tables, unless otherwise specified. For laboratory tests, data from unscheduled visits will be included in listings and summaries of maximum changes from baseline, and the best or worst post-baseline values. For endpoints based on disease status assessment, data from unscheduled assessments will be included in the derivation and analyses of the endpoints.

6.8 Baseline Values

Unless otherwise specified, the baseline value is defined as the value collected at the time closest to, and prior to, the start of study drug administration. Values collected at unscheduled visits prior to the start of the study drug administration will be included in the calculation of baseline values.

6.9 Computing and Coding Standards

Activities will be performed using the following tools:

Table, listing, and figure production	SAS Version 9.2 or higher
Coding	
Adverse Events	MedDRA Version 16.1 or higher
Medical Histories	MedDRA Version 16.1 or higher
Prior and Concomitant Medications	WHODrug Version September 2013
Grading	
AEs	CTCAE Version 4.03
Labs	CTCAE Version 4.03

7 STATISTICAL ANALYSES

7.1 Study Subjects

7.1.1 Disposition of Subjects

The disposition of subjects will include the number and percentage of subjects for the following categories: the number randomized, the number and percentage randomized but not dosed, the number and percentage dosed. These categories will be summarized for each treatment arm and for the two treatment arms combined (total). The percentages will be based on all randomized subjects (ITT analysis set).

An end-of-treatment disposition (still on treatment vs discontinued from treatment) will be provided for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set. The primary reason for treatment discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

An end-of-study disposition (still on study vs discontinued from study) will also be provided for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The primary reason for study discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

A summary of strata as captured in IRT will be presented by treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. A summary table and a listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF will be presented.

7.1.2 Protocol Deviations

Protocol deviations will be categorized as major or minor prior to data release for the interim or final analysis of the primary endpoint. A summary table of the major protocol deviations will be provided by treatment arm and for two arms combined (total) for the ITT analysis set. A listing of major protocol deviations will be provided, broken down by center and treatment arm.

7.1.3 Demographic and Other Baseline Characteristics

Demographic and other baseline variables will be summarized for each treatment arm based on all randomized subjects (ITT analysis set) and for the two treatment arms combined (total). The variables will include age, age group (<65 versus >=65), sex, race, ethnicity, height, weight and the following biomarkers:

- del[17p] (captured from central lab, presence vs absence vs indeterminate)
- del[17p] (captured from local lab, presence vs absence vs indeterminate)
- IgHV status (mutated vs unmutated vs indeterminate)
- CD38 (Positive (≥30%) vs Negative (<30%) vs indeterminate)
- ZAP70 (Positive (>19%) vs Negative (≤19%) vs indeterminate)
- del[17p] and/or TP53 mutation (either or both present vs neither present vs indeterminate; del[17p] will be as captured from central lab)

- TP53 mutation (presence vs absence vs indeterminate)

A separate table will be provided to compare del[17p] values captured from central lab and local lab. Demographic and other baseline variables will also be summarized based on the PP analysis set.

7.1.4 Disease History

Disease history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include diagnosis (CLL or SLL), risk factors used for randomization that are reported on eCRF (high risk cytogenetics [17p deletion: presence vs absence], refractory/early relapse to purine analog based treatment [yes vs no], grade 4 cytopenia(s) [presence vs absence]), years from initial diagnosis to randomization, months from most recent relapse/refractory diagnosis to randomization, Binet/Rai stage at initial diagnosis, type of prior treatment, Binet/Rai stage at baseline, and baseline lymphocyte count.

The durations to be summarized are defined as follows:

- Years from initial diagnosis to randomization will be calculated as (date of randomization – date of initial diagnosis + 1)/365.25.
- Months from most recent relapse/refractory diagnosis to randomization will be calculated as (date of randomization – date of most recent relapse/refractory diagnosis + 1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.5 Prior Therapies

Prior therapies will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include number of prior systemic therapies (summarized as a continuous variable and as a categorical variable) and months from most recent prior systemic therapy to randomization.

The duration to be summarized is defined as follows.

- Months from most recent prior systemic therapy to randomization will be calculated as (date of randomization – stop date of most recent systemic therapy +1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.6 Medical History

Medical history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set.

Medical history will be summarized by system organ class (SOC) and preferred term (PT) using the number and percentage of subjects who had at least one occurrence of an SOC or PT. The summary will be sorted alphabetically in SOC and by decreasing frequency of PT in the duvelisib arm within an SOC.

7.1.7 Prior and Concomitant Medications

Medications will be considered as prior if they stopped before the date of first dose of study drug.

Medications will be considered concomitant if they were taken at any time between the date of first dose of study drug and 30 days after the date of last dose of study drug, inclusive. If the start date or end date of a medication is completely or partially missing, refer to Section 6.2.2 for the algorithm to determine whether a medication is concomitant.

Prior medications and concomitant medications will be summarized separately. Both summaries will be based on the AT analysis set.

Medications will be summarized by ATC level 1, ATC level 2, and preferred drug name for each treatment arm. The summary will be sorted by decreasing frequency in ATC level 1, ATC level 2 and preferred drug name in the duvelisib arm. A subject taking the same drug multiple times will only be counted once.

A listing will be provided for all non-study medications taken on the study. An identifier will be provided to show if a medication is prior or concomitant. Medications that started more than 30 days after the last dose of study drug will be identified as post-treatment.

7.1.8 Exposure to Study Drug

Extent of exposure will be summarized for each treatment arm based on the AT analysis set.

Extent of exposure will be summarized for the following variables:

- Duration (weeks): (date of last dose – date of first dose + 1) divided by 7
- Number of cycles started (continuous and categorical)
- Ofatumumab arm only: For each infusion (1-12), number and percent of subjects receiving the infusion
- Relative dose intensity, defined as 100% x (total dose received)/ (planned cumulative dose for the duration of treatment)
- Number and percentage of subjects with a dose reduction
- duvelisib arm only: Number and percentage of subjects with a dose increase
- Number and percentage of subjects with a dose hold
- Number and percentage of subjects with study drug discontinued

A by-subject listing will be presented for exposure to study drug.

7.2 Efficacy Analyses

For PFS, ORR, lymph node response rate, DOR and other endpoints derived from progression and/or response status, the primary analyses will be based on the endpoints derived from

independent central review by Independent Review Committee (IRC). The endpoints derived from investigator assessment will be used in sensitivity analyses.

All efficacy analyses will be based on the ITT analysis set unless stated otherwise. If analyses are performed on more than one analysis set, the analyses on the ITT analysis set will be considered primary.

For stratified analysis of any efficacy endpoint, the following algorithm will be used to pool strata if there is insufficient information in any stratum (ie, there are <6 subjects, or there is no event for a time-to-event endpoint, or all subject have the same outcome for a binary endpoint in a stratum). (1) Strata will be ranked from the smallest to the largest based on the number of subjects, and if there is a tie, based on the number of events or responses. If there is still a tie, based on the reverse order of strata as determined by the stratification factors and levels (see Section 3.3). (2) The smallest stratum will be compared with the criteria of insufficient information. (3) If there is insufficient information in the smallest stratum, that stratum will be pooled with the smallest of the adjacent strata, which are defined as strata having 2 of the 3 stratification factors being at the same level as the smallest stratum. At the end of the three steps, the pooled stratum will replace the original two contributing strata. If there is still insufficient information in any stratum, the three steps will be repeated with the last pooled stratum assuming the stratum label of the larger of the two contributing strata for the purpose of additional pooling. The resulting pooled strata will be the strata used for the stratified analysis described in this section.

Listings of all efficacy endpoints will be provided.

7.2.1 Analyses of Primary Endpoint

7.2.1.1 Primary Analyses of PFS

The primary analyses of PFS will use PFS based on IRC assessment in the ITT analysis set.

Number of subjects with events, types of events (progression or death before progression), number of subjects censored, number of subjects for each reason of censoring, estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points may also be presented.

A stratified log-rank test (one-sided) will be used to compare PFS of the duvelisib arm against PFS of the ofatumumab arm at the interim and final analyses with the overall one-sided significance level controlled at 0.025. The nominal p-value required for statistical significance at an analysis will be calculated based on the numbers of PFS events at all preceding and current analyses. The HR (duvelisib/ofatumumab) and the corresponding 2-sided 95% CI will be estimated using a stratified Cox proportional hazards model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

PFS will be plotted for each treatment group using the Kaplan-Meier method. The plot will include the number of subjects at risk for each treatment group over time.

7.2.1.2 Sensitivity Analyses of PFS

The following sensitivity analyses will be performed:

- PFS based on investigator assessment: This endpoint will be analyzed using the same methods as the primary analyses described for PFS based on IRC assessment. In addition, the differences between investigator assessment and IRC assessment will be summarized.
- Worst-case sensitivity analysis: Subjects who are alive and have not had documented progression by data cutoff and who are "lost to follow-up" (missing at least one disease assessment right before data cut off) will be treated as censored at their last adequate disease assessment if they are on the control arm (ofatumumab) and treated as having a PFS event at the time of the next scheduled assessment following the last adequate disease assessment if they are on the experimental arm (duvelisib). PFS based on IRC assessment with the above worst-case censoring/event rule will be analyzed using the same methods as the primary analyses for PFS.
- Event-free survival (EFS): This is defined as time from randomization to the first documentation of PD as determined by IRC, start of new anticancer treatment or procedure, or death due to any cause. The event/censoring method is presented in Appendix B. EFS will be analyzed using the same methods as the primary analyses for PFS except that the summary of types of events will include a category of new anticancer treatment or procedure.
- Analysis using the PP analysis set.
- Analysis using the AT analysis set.
- Unstratified analyses: An unstratified log-rank test (one-sided) will be used to compare the two trreatment arms. An unstratified Cox regression (one-sided) will be used to estimate the hazard ratio (duvelisib/ofatumumab) with its 95% confidence interval. Other aspects of the analyses will be the same as the primary analyses of PFS.
- Cox regression with baseline covariates: A stratified Cox regression will be used to test treatment effect on PFS, adjusting for demographic and other baseline characteristics. The strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. A stepwise variable selection will be performed to choose the variables in the Cox regression. Candidate variables are age, gender, race, disease diagnosis (CLL or SLL), years from initial diagnosis, months from most recent relapse/refractory diagnosis, stage at diagnosis, stage at baseline, and number of prior systemic therapies. PFS based on IRC assessment will be used for this analysis.

7.2.1.3 Subgroup analyses of PFS

Subgroup analyses for PFS will be performed using the subgroups specified in Section 6.5. The HR (duvelisib/ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The hazard ratios and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

For selected subgroups, additional analyses of PFS will be performed. Number of subjects with events, types of events (progression or death before progression), number of subjects censored, number of subjects for each reason of censoring, estimates and 2-sided 95% confidence intervals for the 25th percentile, median, and 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points may also be presented. PFS will be plotted for each treatment group using the Kaplan-Meier method. The plot will include the number of subjects at risk for each treatment group over time.

7.2.2 Analyses of Secondary Efficacy Endpoints

Of the secondary endpoints, ORR and OS are designated as key secondary efficacy endpoints. These endpoints will be tested for statistical significance under an overall one-sided significance level of 0.025 if and only if the primary endpoint is significant. Details of multiplicity adjustment are provided in Section 6.3.

The analyses of secondary efficacy endpoints, Lymph Node Response Rate and Hematologic Improvement Rate, will include hypothesis testing, but no statistical significance will be claimed. The analyses of other secondary efficacy endpoint, Duration of Response (DOR), will be nominal only.

7.2.2.1 Overall Response Rate (ORR)

The primary analyses of ORR will use ORR based on IRC assessment in the ITT analysis set.

ORR will be derived from best overall response (BOR), which is defined as the best time point response that a subject achieves during the course of the study, with the response ranked according to the following order (from best to worst): CR>CRi>PR>PRwL>SD>PD (CRi applies to CLL only). In addition, the IRC may assign a timepoint response of unknown (UNK) due to missing, incomplete, or inadequate data; at baseline, the IRC may assign a timepoint response of no evidence of disease (NED) if both radiological and clinical data indicate no disease involvement; for a post-baseline timepoint, the IRC may also assign a timepoint response of not evaluable (NE) if no target lesions were identified at baseline and the radiological and clinical data at the post-baseline timepoint do not support the disease response of PD or UNK; these categories will be classified as OTHER in the summary.

The estimated ORR (percent of subjects with a BOR of CR, CRi, PR or PRwL) and a 2-sided 95% CI will be provided for each treatment arm. Number and percent of subjects with BOR in each of the response categories (CR, CRi, PR, PRwL, SD, PD, OTHER) will also be presented by treatment arm. All subjects in the analysis set will be included in the denominator in the calculation of the percentage for each response category or ORR.

ORR will be analyzed using the Cochran-Mantel-Haenszel (CMH) test (one-sided, see Appendix C) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. The odds ratio and its 95% CI will be estimated.

The following sensitivity analyses will be performed:

- ORR based on investigator assessment
- Overall Confirmed Response Rate (OCRR), with overall confirmed response (based on independent review) defined as best confirmed response (time between response and confirmation must be ≥8 weeks in duration) of CR, CRi, PR, or PRwL, according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Overall Response Rate without PRwL: with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR)
- Overall Confirmed Response Rate without PRwL
- Overall Response Rate in subset of subjects with baseline assessment other than UNK and NED. This analysis will be performed if there are at least 5% subjects with baseline assessment of UNK or NED
- Overall Response Rate in subset of subjects who are on treatment for 60 days or longer
- Analysis using the PP analysis set
- Analysis using the AT analysis set

Subgroup analyses will be performed for ORR using the subgroups specified in Section 6.5.

The odds ratio (duvelisib/ofatumumab) with its 95% CI will be calculated for each subgroup based on unstratified test. The odds ratios and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

For selected subgroups, the estimated ORR and its 2-sided 95% CI will be provided for each treatment arm; number and percent of subjects with BOR in each of the response categories (CR, CRi, PR, PRwL, SD, PD, OTHER) will also be presented by treatment arm.

7.2.2.2 Overall Survival (OS)

The primary analyses of OS will be based on the ITT analysis set.

Subjects without documentation of death at the time of the data cutoff for analysis will be censored at the date the subject was last known to be alive, or the data cutoff date, whichever is earlier. A stratified one-sided log-rank test will be used to compare OS between the 2 treatment groups. The HR along with the 95% CI will be estimated using a stratified Cox model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. A Kaplan-Meier plot for OS will be presented by treatment group. Estimates and 95% confidence intervals for the

25th percentile, median, and 75th percentile for OS will be presented by treatment group (if estimable). Probabilities of survival at selected time points may also be presented.

A majority of subjects can be expected to take subsequent anticancer therapy, especially after disease progression. To adjust for the effects of subsequent therapy, the following sensitivity analysis may be performed:

1. Primary analyses except additional censoring at the start date of subsequent therapy
2. Rank preserving structural failure time (RPSFT) analysis
3. Inverse probability of censoring weighted (IPCW) analysis

The following sensitivity analyses will be performed:

- Analysis using the PP analysis set.
- Analysis using the AT analysis set.
- Unstratified analyses: An unstratified log-rank test (one-sided) will be used to compare the two treatment arms. An unstratified Cox regression (one-sided) will be used to estimate the hazard ratio (duvelisib /ofatumumab) with its 95% confidence interval.

Subgroup analyses for OS will be performed using the subgroups specified in section 6.5. The HR (duvelisib /ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The HRs and their 95% CIs will be displayed for all subgroups graphically in a forest plot.

OS follow-up time, defined as time from randomization to the date of last procedure or assessment (including telephone interview), will be summarized by treatment arm and for the two treatment arms combined (total) using the reverse Kaplan-Meier method, which applies Kaplan-Meier method to the OS data with the same event/censoring time as the primary OS analyses while reversing event/censoring indicator. The 25th percentile, median, and 75th percentile of OS follow-up time will be presented.

7.2.2.3 Lymph Node Response Rate

The primary analyses of lymph node response rate will be based on IRC assessment in the ITT analysis set. The number of subjects with lymph node response, estimated lymph node response rate, and a 2-sided 95% CI will be provided for each treatment arm.

Lymph node response rate will be analyzed using the CMH test (one-sided) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

A sensitivity analysis will be performed using lymph node response rate based on investigator measurements.

7.2.2.4 Hematologic Improvement Rate

A subject with a hematologic improvement is one who consistently met the criteria of an improvement in neutrophil count, hemoglobin or platelet count for a period of at least 60 days during which the subject did not have a transfusion or exogenous cytokines. Missing data or missing scheduled assessments during a potential 60-day hematologic improvement episode will result in categorization of not having a hematologic improvement.

The number of subjects with hematologic improvement, estimated hematologic improvement rate, and a 2-sided 95% CI will be provided for each treatment arm.

Hematologic improvement rate will be compared between treatment groups using one-sided CMH test. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

7.2.2.5 Duration of Response (DOR)

DOR will be presented using the Kaplan-Meier method for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL) as determined by IRC assessment. The censoring method will be the same as that for the primary endpoint (see Appendix A). The Kaplan-Meier plot for DOR will be presented by treatment group. Estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for DOR will be presented by treatment group (if estimable). No treatment comparison will be performed.

7.2.3 Analyses of Exploratory Efficacy Endpoints

7.2.3.1 Improvement in Disease-Related Symptoms

Disease-related symptoms of fever, weight loss, and drenching night sweats will be assessed as present or absent at each timepoint. Improvement in each of the three disease-related symptoms from baseline will be analyzed separately. The summary will display by treatment arm the proportion of subjects with an improvement while on treatment. Descriptive analysis of improvement in these disease-related symptoms will also be presented over the course of the study (eg, for each scheduled assessment) by treatment arm.

The disease-related symptom of fatigue will be measured by the ECOG performance status. Summary statistics will be provided by treatment arm for the ECOG score and the change from baseline at each timepoint.

7.2.3.2 Minimal Residual Disease

MRD status for subjects with a response of CR or CRi will be determined using flow cytometry from a central lab. MRD will be classified as positive or negative. The estimated rates of subjects with negative MRD and their 95% confidence intervals in subjects with a response of CR or CRi will be provided for each treatment arm.

7.2.3.3 Time To Response (TTR)

TTR will be presented for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL). Summary statistics and a 2-sided confidence interval for median TTR will be provided by treatment arm. No treatment comparison will be performed.

7.3 Safety Analyses

All safety analyses will be performed using the AT analysis set.

7.3.1 Adverse Events

Adverse events will be coded using MedDRA Version 16.1 or higher. The Grade of the AE will be assessed according to the NCI-CTCAE Version 4.03. If an AE is not included in the NCI-CTCAE Version 4.03, the Grade of the AE will be assessed according to the protocol, Section ***8.2.1.2***.

The summary of AEs will be focused on treatment-emergent AEs. A treatment-emergent AE (TEAE) is defined as any AE that emerges or worsens in the period from the first dose of study treatment to 30 days after the last dose of study treatment. The onset date of an AE will be compared to the first dose date and the last dose date plus 30 days to determine whether the AE is treatment-emergent or not. If the onset date of an AE is completely or partially missing, refer to Section 6.2.1 for the algorithm to determine whether an AE is treatment emergent.

TEAEs will be summarized for each treatment arm by MedDRA system organ class (SOC) and preferred term (PT), or PT only. For summary tables by SOC and PT, SOC will be sorted alphabetically and PT will be sorted by decreasing frequency in the duvelisib arm within each SOC. For summary tables by PT only, PT will be sorted by decreasing frequency in the duvelisib arm.

If multiple TEAEs of the same PT occur within a subject, only the maximum grade observed for this PT will be used in summary of TEAEs by grade, the subject will be counted only once in the number of subjects for this PT and only once for the number of subjects for the SOC to which this PT belongs.

An overview TEAE summary table will be provided, which will include the number of subjects with AEs in selected categories. In addition, TEAEs will be summarized for the following categories, and will be tabulated by SOC and PT, unless otherwise specified.

- Treatment-emergent AEs (All Causalities)
- Treatment-emergent AEs (Treatment-Related)
- Treatment-emergent AEs (All Causalities, by maximum grade)
- Treatment-emergent AEs (Treatment-Related, by maximum grade)
- Grade 3 or higher treatment-emergent AEs (All Causalities)
- Grade 3 or higher treatment-emergent AEs (Treatment-Related)

- Treatment-emergent SAE (All Causalities)
- Treatment-emergent SAE (Treatment-Related)
- Treatment-emergent AEs resulting in discontinuation of study drug (All Causalities)
- Treatment-emergent AEs resulting in dose hold (All Causalities)
- Treatment-emergent AEs resulting in dose reduction (All Causalities)
- Treatment-emergent AEs resulting in dose hold or reduction (All Causalities)
- Treatment-emergent AEs resulting in discontinuation of study drug (Treatment-Related)
- Treatment-emergent AEs resulting in death (All Causalities)
- Treatment-emergent AEs resulting in death (Treatment-Related)
- Treatment-emergent AEs by PT (All Causalities)
- Treatment-emergent AEs by PT (AEs with onset date within 24 weeks after the first dose of study treatment; All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (AEs with onset date within 24 weeks after the first dose of study treatment; All Causalities)
- Treatment-emergent SAEs by PT (All Causalities)
- Treatment-emergent SAEs by PT (Treatment-Related)
- Treatment-emergent AEs reported in ≥5% of subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
- Grade 3 or higher treatment-emergent AEs reported in ≥5% of subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)

A by-subject listing of the following AE categories will be presented.

- All AEs (TEAEs will be flagged)
- All SAEs (TEAEs will be flagged)
- Treatment-emergent AEs resulting in dose hold
- Treatment-emergent AEs resulting in dose reduction
- Treatment-emergent AEs resulting in discontinuation of study drug
- Treatment-emergent AEs resulting in death

Treatment-emergent AEs (All Causalities) and Treatment-emergent AEs (Treatment-Related) will also be tabulated in each subgroup specified in Section 6.5.

7.3.2 Laboratory Data

Laboratory tests will be reported separately for hematology and blood chemistry.

For the purposes of presentation in both tables and listings, the following laboratory test results will be converted to the International System of Units (SI) before presentation: sodium, potassium, chloride, bicarbonate (or CO_2), albumin, total protein, creatinine, uric acid, calcium, phosphorus, magnesium, glucose, total and direct bilirubin, and alkaline phosphatase, red blood cell (RBC) count, hemoglobin, hematocrit, platelets, white blood cell count with 5-part differential performed manually or by flow cytometry (lymphocytes, neutrophils, monocytes, basophils, and eosinophils), etc.

Lab tests performed after the start of the first dose of study treatment, up to 30 days from the last dose, will be considered treatment-emergent. Only treatment-emergent lab tests will be included in the analyses in this section.

If a laboratory test value is reported using a non-numeric qualifier (e.g., less than [<] a certain value, or greater than [>] a certain value), the given numeric value will be used in the summary statistics, ignoring the non-numeric qualifier.

For laboratory tests with NCI-CTCAE grades, a shift table from baseline grade to the maximum post-baseline grade will be provided. Laboratory tests with bi-directional grades (e.g., Hyperglycemia and Hypoglycemia) will be presented separately for each direction within the shift table.

Listings will be provided for all laboratory test results and for laboratory test results grade 3 and higher. A listing of subjects with ALT or AST >3xULN with simultaneous total bilirubin >2xULN will be presented, where ULN stands for upper limit of normal. The elevations of ALT/AST and total bilirubin must occur within 2 days of each other.

7.3.3 Vital Signs

The actual values of vital sign parameters, including temperature, heart rate, weight and systolic and diastolic blood pressure, will be presented in a by-subject listing.

7.3.4 Electrocardiogram (ECG)

A by-subject listing will be presented for baseline ECGs and post-baseline unscheduled ECGs (if any).

7.3.5 Concomitant Medications and Procedures

Please refer to Section 7.1.7 for the definition and summary of concomitant medications.

Concomitant procedures will not be summarized. A by-subject listing will be presented.

7.3.6 Deaths

On-treatment deaths are defined as deaths that occur within 30 days of the last dose of study medication. Deaths on treatment and causes for deaths occurring on-treatment, and deaths during follow-up will be summarized.

A by-subject listing of deaths on study will be presented.

7.4 Quality of Life Instruments

Analyses of the QoL instruments (EQ-5D and FACIT-F) will be performed on the ITT analysis set. Additional analyses may be considered as needed.

7.4.1 EQ-5D

The EQ-5D contains a descriptive system with one response for each of 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). The first response is coded as 1 (indicating no problems), the second response is coded as a 2 (indicating some problems), and the third response is coded as a 3 (indicating extreme problems). Ambiguous responses (eg, more than one response in a dimension) are treated as missing values. The EQ-5D also contains a Visual Analog Scale (VAS) for health state ranging from 0 to 100, where 0 represents worst imaginable health state and 100 represents the best imaginable health state.

For each dimension of the descriptive system, the number and percentage of subjects with no problems, some problems, and extreme problems will be reported at each visit.

Summary statistics for the VAS of health state and the change from baseline will be reported for each visit.

7.4.2 FACIT-F

The FACIT-F contains 5 subscales, Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, Functional Well-Being, and Fatigue (Additional Concerns on eCRF). Each subscale contains multiple items, and each item has a response of 0 to 4. Scores will be computed for each subscale, as well as for the FACIT-F Trial Outcome Index, FACT-G total score, and the FACIT-F total score.

The Physical Well-Being subscale contains 7 items. To compute the subscale score, compute the item score by subtracting each item response from 4 (item score = 4 – item response), sum the item scores, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Physical Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Social/Family Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Social/Family Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Emotional Well-Being subscale contains 6 items. To compute the subscale score, compute the item score first. For items 1, 3, 4, 5, and 6, subtract each item response from 4 (item score = 4 – item response). For item 2, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 6 and divide by the number of items answered to compute the subscale score. At least 3 questions must be answered. If fewer than 3 questions are answered, then the subscale score will be missing. The Emotional Well-Being score ranges from 0 to 24 with higher scores representing better quality of life.

The Functional Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Functional Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Fatigue subscale contains 13 items. To compute the subscale score, compute the item score first. For all items except An5 and An7, subtract each item response from 4 (item score = 4 – item response). For items An5 and An7, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 13 and divide by the number of items answered to compute the subscale score. At least 7 questions must be answered. If fewer than 7 questions are answered, then the subscale score will be missing. The Fatigue score ranges from 0 to 52 with higher scores representing better quality of life.

To derive the FACIT-F Trial Outcome Index, sum the subscale scores for Physical Well-Being, Functional Well-Being, and Fatigue. If any of these subscale scores is missing, the score for the FACIT-F Trial Outcome Index will be missing. The score ranges from 0 to 108.

To derive the FACT-G score, sum the subscale scores for Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, and Functional Well-Being. If any of these subscale scores is missing, the FACT-G score will be missing. The score ranges from 0 to 108.

To derive the FACIT-F total score, sum all 5 subscale scores. If any subscale score is missing, the FACIT-F total score will be missing. The score ranges from 0 to 160.

For each of the five subscale scores and for each of the FACIT-F Trial Outcome Index, the FACT-G score, and the FACIT-F total score, summary statistics for the actual values and changes from Baseline will be reported at each visit.

7.5 Pharmacokinetic Analyses

Plasma samples will be analyzed for duvelisib and IPI-656 (metabolite) concentrations using a validated high performance liquid chromatography-tandem mass spectrometry (LC/MS/MS) method. A listing of the duvelisib and IPI-656 concentrations will be provided; the sample collection date and time, and the date, time and amount of the preceding duvelisib dose administration will be included in the listing.

The PK data collected will be analyzed by standard population PK methods, using appropriate software. Analysis of exposure-response relationships for efficacy and safety endpoints will be

conducted. If there is only a limited amount of plasma concentration data from this study, the data may be pooled with the results of other studies to perform the population PK and exposure-response analyses. Further details on these analyses will be outlined in a separate analysis plan. Results of the population PK and exposure-response analysis will be summarized in a separate technical report.

7.6 Biomarker Analyses

The following exploratory analyses may or may not be presented formally. If presented, they will be limited to biomarkers for which data from a sufficient number of subjects are available. Further details on these analyses may be outlined in a separate analysis plan, and results may be summarized in a separate technical report.

7.6.1 Pharmacodynamic and Potentially Predictive Biomarkers

The relationship between tumor genomics, gene expression, protein expression, and clinical endpoints may be explored as follows:

- Evaluation of serum biomarkers, blood immunophenotype and tumor biomarkers for relationships with duvelisib clinical activity and safety
- Evaluation of tumor genomic features (eg, cytogenetics, FISH, DNA sequence variation, DNA copy number variation, and/or RNA expression) for relationships with duvelisib clinical activity
- Evaluation of the relationship between serum and/or tumor biomarkers and disease progression of duvelisib and ofatumumab

7.6.2 Pharmacogenomics

DNA may be extracted from the pharmacogenomics sample in order to evaluate a subject's germline DNA. This sample may be used to explore the following:

- Evaluate the relationship between germline DNA sequence variations and PK of duvelisib
- Evaluate germline DNA as a control to verify that tumor sequence variations are somatic mutations

8 CHANGES IN PLANNED ANALYSES

8.1 Changes in Planned Analyses from Protocol Amendment 1 to SAP V1.0

There are no changes from Protocol Amendment 1. More details are provided in the SAP.

8.2 Changes in Planned Analyses from SAP V1.0 to SAP V2.0

The main changes from SAP V1.0 to SAP V2.0 are described below.

1. Section 3.5.2 Clarification

   Old (V1.0): The second analysis of OS will be performed at the PFS final analysis, which is expected to happen after 185 PFS events have occurred, or approximately 27.5 months after the first subject is randomized.

   New (V2.0): The second interim analysis of OS will be performed at the PFS final analysis, which is to happen after 185 PFS events have occurred. If the planned final analysis of PFS is not performed due to early stop for efficacy, then the second interim analysis of OS will be conducted approximately one year after the first interim analysis of OS.

2. New censoring method for event-free survival were specified.
3. Prior irradiation and prior surgery were removed from prior therapy summary.

The following additional analyses/summaries have been added to this version of the SAP:

1. Analysis of ORR of subjects who are on treatment for 60 days or longer
2. Time to OS follow up
3. Time to response
4. Treatment-emergent AEs reported in ≥5% of CLL/SLL subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
5. Grade 3 or higher treatment-emergent AEs reported in ≥5% of CLL/SLL subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
6. Treatment-emergent AEs by PT (All Causalities, AE onset date within 24 weeks after the first dose of study treatment)
7. Grade 3 or higher treatment-emergent AEs by PT (All Causalities, AEs with onset date within 24 weeks after the first dose of study treatment)
8. Treatment-emergent AEs resulting in discontinuation of study drug (Treatment-Related)
9. Treatment-emergent AEs resulting in death (Treatment-Related)
10. Treatment-emergent SAEs by PT (All Causalities).
11. Treatment-emergent SAEs by PT (Treatment-Related)
12. Deaths on treatment
13. Deaths during follow-up

14. A by-subject listing of treatment-emergent AEs resulting in dose hold
15. A by-subject listing of treatment-emergent AEs resulting in dose reduction
16. A by-subject listing of deaths on study
17. A summary table and a listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF
18. Comparision of del[17p] values captured from central lab and local lab
19. The following biomarkers will be summarized as baseline characteristics:
    - del[17p] (captured from central lab, presence vs absence vs indeterminate)
    - del[17p] (captured from local lab, presence vs absence vs indeterminate)
    - del[17p] and/or TP53 mutation (del[17p] and/or TP53 mutation, del[17p] captured from central lab, either or both present vs neither present vs indeterminate)
    - del[17p] and TP53 mutation (del[17p] and TP53 mutation, del[17p] captured from central lab, both present vs either vs neither present vs indeterminate)
    - TP53 mutation (presence vs absence vs indeterminate)
    - IgHV status (mutated vs unmutated vs indeterminate)
    - CD38 (Positive (≥30%) vs Negative vs indeterminate)
    - ZAP70 (Positive (>19%) vs Negative (<=19%) vs indeterminate)
20. Subgroup analyses of PFS, OS and ORR based on the following factors:
    - Time from last dose of most recent prior anti-cancer therapy to randomization <12 months: yes vs no
    - del[17p] (captured from central lab) or TP53 mutation (either or both present vs neither present)

8.3 Changes in Planned Analyses from SAP V2.0 to SAP V3.0

The main changes from SAP V2.0 to SAP V3.0 are described below.

1. In Section 3.5.2, the estimated numbers of OS events at each of OS interim analyses and final analysis are provided.
2. In Section 6.3, Lymph Node Response Rate and Hematological Improvement Rate are removed from the list of key secondary efficacy endpoints and removed from the hierarchy of inferential testing in secondary endpoints.
3. In Section 7.2.1.3, for selected subgroups, additional analyses of PFS are added. These include the estimate and 95% CI for the median by treatment group, and the Kaplan-Meier plot by treatment group.
4. In Section 7.2.2.1, for selected subgroups, summaries of BOR, estimated ORR and its 2-sided 95% CI for each treatment group are added.

9 REFERENCES

1 O'Brien PC, Fleming TR. *A multiple testing procedure for clinical trials.* Biometrics 1979; 35:549-556.

2 Robins JM. *Information recovery and bias adjustment in proportional hazard regression analysis of randomized trials using surrogate markers*. Proceedings of the Biopharmaceutical Section, American Statistical Association 1993; pp. 24-33.

3 Robins JM, Finkelstein D. *Correcting for Non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests*. Biometrics, 2000; 56(3):779-788.

4 Robins JM, Tsiatis, AA. *Correcting for noncompliance in randomized trials using rank preserving structural failure time models.* Communications in Statistics, 1991; 20, 2609-2631.

10 APPENDICES

10.1 Appendix A: Primary PFS Event/Censoring Method

Censoring of PFS will be performed as detailed in the table below.

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression or death before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death	Date of death	Event
Documented progression or death following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment or procedure started before documented progression	Date of last adequate disease status assessment	Censored

Note: Disease status assessment includes CT scans (chest, abdomen and pelvis), bone marrow aspirate and/or biopsy (may not be required of all subjects at all scheduled disease status assessments), CBC and differential count, focused physical examination, disease related constitutional symptoms for disease assessment, and ECOG performance status. An adequate baseline disease status assessment is any baseline disease status assessment that include WBC counts and CT scans. An adequate post-baseline disease status assessment is any disease status assessment for which a disease status (eg, CR, CRi, PR, PRwL, SD, and PD) is arrived per protocol-defined criteria by the IRC (for IRC assessment) or investigator (for investigator assessment).

10.2 Appendix B: Event-Free Survival (EFS) Event/Censoring Method

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death or new anticancer treatment/procedure occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression, death or new anticancer treatment/procedure started before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression, and no new anticancer treatment/ procedure started before documented progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death, and no new anticancer treatment/procedure started before death	Date of death	Event
Documented progression, death or new anticancer treatment/procedure following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment/procedure started before documented progression with ≤1 missing scheduled disease status assessment prior to new anticancer treatment/procedure	Start date of new anticancer treatment/procedure	Event

10.3 Appendix C: One-Sided Cochran–Mantel–Haenszel (CMH) Test

The Cochran–Mantel–Haenszel (CMH) test compares binary responses of two treatment groups, adjusting for stratification factors. In the CMH test, the data are arranged in a series of associated 2 × 2 contingency tables, the null hypothesis is that the observed response is independent of the treatment used in any 2 × 2 contingency table.

Let O_{hij} be the observed frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h and E_{hij} be the expected frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h, where

$$E_{hij} = \frac{(O_{hi1} + O_{hi2}) \cdot (O_{h1j} + O_{h2j})}{O_{h11} + O_{h12} + O_{h21} + O_{h22}}$$

Also, let P_{hij} be the probability of outcome being j given treatment being i (i=1 or 2, j= 1 or 2) in stratum h. To test the following hypothesis,

$$H_0 : P_{h11} = P_{h21} \text{ for all } h \in \{1, \ldots, H\}$$
versus
$$H_1 : P_{h11} > P_{h21} \text{ for at least one } h \in \{1, \ldots, H\}$$

The one-sided Cochran-Mantel-Haenszel Test Statistics is constructed as:

$$Z_{CMH} = \frac{\sum_{h=1}^{H} (O_{h11} - E_{h11})}{\sqrt{V_{11}}} \sim Z$$

Where

$$V_{11} = Var(\sum_{h=1}^{H} (O_{h11} - E_{h11})) = \sum_{h=1}^{H} \frac{E_{h11} \bullet E_{h22}}{O_{h11} + O_{h12} + O_{h21} + O_{h22} - 1}$$

The test statistic Z_{CMH} will be compared with standard normal distribution to obtain p-value of the CMH test.

STATISTICAL ANALYSIS PLAN

Protocol IPI-145-07

A Phase 3 Study of IPI-145 versus Ofatumumab in Subjects with Relapsed or Refractory Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma

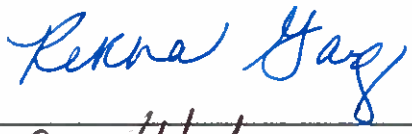
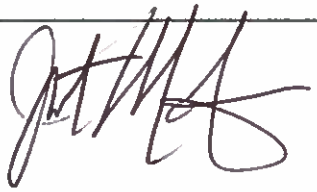
Protocol Version:	**Amendment 2**
Type of Analysis Plan:	**Interim and Final Analyses**
Version:	**2.0**
Date:	**24 Nov 2015**
Author:	**Shijie Tang, PhD**

STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

The undersigned has developed this statistical analysis plan (SAP):

Name/Title	Signature	Date
Shijie Tang, PhD Associate Director Biostatistics		24 Nov 2015

The undersigned have reviewed this SAP and approve it in its entirety:

Name/Title	Signature	Date
Virginia Kelly, MD Senior Director Clinical Development		02 Dec 2015
Rekha Garg, MD, MS Vice President Safety and Risk Management		02 Dec 2015
Justin McLaughlin, MA Director Medical Writing		24 Nov 2015
Lixin Han, PhD Director Biostatistics		02 Dec 2015

DOCUMENT HISTORY

Version	Date	Author(s)	Brief Summary of Changes
1.0	02 Oct 2014	Shijie Tang	Original
2.0	24 Nov 2015	Shijie Tang	Clarified the timing of the 2nd interim analysis of overall survival; added biomarkers to baseline characteristics; other changes. Refer to Section 8.2 for Details.

Table of Contents

LIST OF ABBREIVATIONS

Abbreviation	Description
ADaM	Analysis Data Model
AE	Adverse Event
ANCOVA	Analysis of Covariance
AT	All-Treated
ALT	Alanine Aminotransferase
ANC	Absolute Neutrophil Count
AST	Aspartate Aminotransferase
ATC	Anatomical Therapeutic Chemical
BID	Twice a day
BOR	Best Overall Response
BUN	Blood Urea Nitrogen
CI	Confidence Interval
CLL	Chronic Lymphocytic Leukemia
CMH	Cochran–Mantel–Haenszel
CO2	Bicarbonate
CR	Complete Response
CRi	Complete Response with Incomplete Marrow Recovery
CS	Abnormal and Clinically Significant
CTCAE	Common Terminology Criteria for Adverse Events
DOR	Duration of Response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form
eDISH	Evaluation of Drug Induced Serious Hepatotoxicity
EQ-5D	European Quality of Life-5 Dimensions
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue
FACT-G	The Functional Assessment of Cancer Therapy - General

Abbreviation	Description
HR	Hazard Ratio
IDMC	Independent Data Monitoring Committee
IPI-145	(S)-3-(1-(9H-purin-6-ylamino)ethyl)-8-chloro-2-phenylisoquinolin-1(2H)-one
IPCW	Inverse Probability of Censoring Weighted
IRC	Independent Review Committee
IRT	Interactive Response Technology
ITT	Intent-To-Treat
IV	Intravenous
IWCLL	International Workshop on Chronic Lymphocytic Leukemia
IWG	International Working Group
LC/MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LDH	Lactate Dehydrogenase
LFT	Liver Function Test
MedDRA	Medical Dictionary for Regulatory Activities
MEOI	Medical Events of interest
MMRM	Mixed Effect Repeated Measures
MRD	Minimal Residual Disease
NCI	National Cancer Institute
NCS	Abnormal but Not Clinically Significant
NED	No Evidence of Disease
ORR	Overall Response Rate
OS	Overall Survival
PD	Progressive Disease
PFS	Progression-Free Survival
PK	Pharmacokinetics
PP	Per-Protocol
PR	Partial Response
PRwL	PR with Lymphocytosis

Abbreviation	Description
PT	Preferred Term
QoL	Quality of Life
RBC	Red Blood Cell
RPSFT	Rank Preserving Structural Failure Time
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
SI	Standard International System of Units
SLL	Small Lymphocytic Lymphoma
SOC	System Organ Class
SPD	Sum of Products
TEAE	Treatment Emergent Adverse Event
TTR	Time to Response
UNK	Unknown
VAS	Visual Analog Scale
WHO	World Health Organization
WHO-DD	World Health Organization Drug Dictionary

1 INTRODUCTION

This is the statistical analysis plan (SAP) for study IPI-145-07, *A phase 3 study of IPI-145 versus ofatumumab in subjects with relapsed or refractory chronic lymphocytic leukemia/small lymphocytic lymphoma*. This SAP is prepared according to Amendment 2 of the protocol, dated 02 March 2015.

2 STUDY OBJECTIVES

2.1 Primary Objective

The primary objective of this trial is to examine the efficacy of duvelisib (IPI-145) monotherapy versus ofatumumab monotherapy in subjects with relapsed or refractory Chronic Lymphocytic Leukemia (CLL) or Small Lymphocytic Lymphoma (SLL).

2.2 Secondary Objectives

- To determine the safety of duvelisib in subjects with CLL or SLL
- To evaluate the pharmacokinetics (PK) of duvelisib and, if applicable, its metabolite(s)

2.3 Exploratory Objectives

- To evaluate the health-related quality of life (QoL) of subjects
- To evaluate pharmacodynamic biomarkers of duvelisib
- To evaluate biomarkers that may predict duvelisib clinical activity and/or safety
- To evaluate mechanisms of resistance in subjects who exhibit disease progression while being treated with duvelisib or ofatumumab
- To evaluate genomic features of tumors predictive of response in subjects treated with duvelisib or ofatumumab

3 STUDY DESIGN

3.1 Overview

Study IPI-145-07 is a two-arm, randomized (1:1 ratio), open-label, phase 3, superiority trial designed to evaluate the efficacy and safety of duvelisib compared to ofatumumab.

Subjects who meet all the eligibility criteria at Screening will return to the clinic on Day 1 to receive their first dose of study drug (randomized to either duvelisib or ofatumumab). The first treatment cycle for each treatment arm will be 3 weeks (21±2 days). Subsequent treatment cycles will be 4 weeks (28±4 days).

Subjects randomized to duvelisib will be given a starting dose of 25 mg duvelisib administered orally twice daily (BID) initially in a 21-day treatment cycle followed by 28-day treatment cycles for up to 18 cycles or until disease progression or unacceptable toxicity (whichever comes first). After completing approximately 18 cycles of treatment, subjects may receive additional cycles of duvelisib for up to 3 years (39 cycles of total treatment) if they have documented evidence of response and disease requiring continued treatment according to the modified IWCLL/revised IWG criteria.

Subjects randomized to ofatumumab will receive treatment consistent with the approved product labeling: 8 weekly infusions, starting with an initial IV dose of 300 mg ofatumumab on Day 1 followed by seven weekly doses of 2000 mg. Thereafter, subjects will receive 2000 mg ofatumumab once every month for four months or until disease progression or unacceptable toxicity (whichever comes first), as outlined in the schedule of assessments. Administration of ofatumumab will not exceed the 12 doses (within 7 cycles) as described in the prescribing information.

Subjects will be followed for survival for up to 6 years (or other duration as specified by the protocol) from randomization or until death. Follow-up visits will occur every 6 months which can be conducted through a telephone interview.

3.2 Sample Size Consideration

This study employs a randomized, open-label, parallel design to assess the potential superiority of duvelisib treatment over ofatumumab treatment on progression-free survival (PFS) in CLL or SLL subjects.

The sample size was determined to provide sufficient power to test the primary endpoint of PFS in a group sequential design with one planned interim analysis.

Assuming an exponential distribution for PFS, a total of 185 PFS events (deaths or progressions) will provide approximately 93% power to detect a hazard ratio of 0.6 (median PFS of 9 months in the ofatumumab group versus 15 months in the duvelisib group) using a one-sided log-rank test at a 2.5% overall significance level (ie, alpha of 0.025). The study design employs the Lan-DeMets spending function for O'Brien-Fleming boundary as the alpha spending function and the Hwang-Shih-DeCani gamma (-4) spending function as the beta spending function. A total of 300

subjects will be randomized in a 1:1 ratio to receive either ofatumumab or duvelisib. Assuming a peak accrual rate of 30 subjects per month with 11 months for accrual ramp-up and a 4% cumulative dropout rate per year, the enrollment would complete in 16 months, with the final analysis for PFS estimated to occur approximately 27.5 months from the randomization of the first subject.

3.3 Randomization

Once a subject has met all entry criteria, the Interactive Response Technology (IRT) will be used to generate a distinct subject identifier. If a subject discontinues from the study, the subject identifier will not be reused and the subject will not be allowed to re-enter the study.

Eligible subjects will be randomized via IRT in a 1:1 ratio to one of two treatment arms:

- Arm 1: duvelisib 25 mg BID
- Arm 2: Ofatumumab

In order to ensure subject balance between treatment groups, study subjects will be stratified by the following:

- High risk cytogenetics (presence vs absence of del[17p])
- Refractory/early relapse to purine analog based treatment (progression <12 months after fludarabine/pentostatin: yes vs no)
- Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)

First dose is to occur within 7 days of randomization after all screening assessments have been completed.

3.4 Blinding

This study is open-label and Investigators, site staff, Sponsor and Sponsor designees will have access to treatment assignments for individual subjects. To reduce bias, a blinded independent central review of disease status will be conducted, which will be blinded to individual subject treatment assignments. Investigators, site staff, Sponsor and Sponsor designees will only have access to blinded (pooled) aggregate study data, except for a select number of Sponsor staff, who may review aggregate SAEs and Medical Events of Interest (MEOI) data by treatment groups.

3.5 Planned Analyses

Interim and final analyses are planned for PFS and OS. For an interim analysis, the actual p-value boundaries will be calculated based on the actual number of PFS and OS events at the analysis. For the final analysis, the actual p-value boundary for efficacy will be calculated based on the actual number of PFS and OS events at the interim analysis (or analyses) and the final analysis.

The Independent Data Monitoring Committee (IDMC) will review PFS, OS and other efficacy data at the time of PFS interim analysis. The list of IDMC deliverables, including tables and listings, is included in the IDMC charter. These deliverables will be based on the methods described in this SAP.

3.5.1 Planned Analyses for PFS

There will be two planned analyses for PFS, one interim analysis and one final analysis. After approximately 50% of the planned PFS events (ie, 93 PFS events) have been observed based on the blinded, independent, central review, an interim analysis of efficacy will be performed.

At the interim analysis, PFS will be tested at the alpha level based on the Lan-DeMets alpha spending function for the O'Brien-Fleming boundary with the opportunity to stop the study (other than survival follow-up) for overwhelming evidence of efficacy. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for efficacy is one-sided p-value of 0.0015 (corresponding approximately to a HR of 0.540). In the meantime, PFS will also be tested at the alpha level for futility based on the alpha and beta spending functions specified for the study design. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for futility is one-sided p-value of 0.4791 (corresponding approximately to a HR of 0.990). The futility boundary of this study is non-binding, meaning that the Type I error is properly controlled even if the study is continued after the futility boundary for PFS is crossed at the interim analysis. Under the protocol design assumptions (eg, the study is fully enrolled in 16 months), the interim analysis is expected to occur approximately 17 months after the first subject is randomized.

If the study is not stopped at the interim analysis, the final analysis will be performed when approximately 185 PFS events have occurred in the study. If the interim analysis and the final analysis occur at exactly 93 and 185 PFS events, the criterion of statistical significance (ie, boundary for efficacy) is one-sided p-value of 0.0245 (corresponding approximately to a HR of 0.748).

3.5.2 Planned Analyses for OS

There will be three planned analyses for OS, two interim analyses and one final analysis. The first interim analysis of OS will be at the planned PFS interim analysis, which is to happen after 93 PFS events have occurred. The second interim analysis of OS will be performed at the PFS final analysis, which is to happen after 185 PFS events have occurred. If the planned final analysis of PFS is not performed due to early stop for efficacy, then the second interim analysis of OS will be conducted approximately one year after the first interim analysis of OS. Since all subjects will be followed for survival for 6 years, the final analysis of OS will take place at the end of follow-up of all subjects, or approximately 6 years from the last subject being randomized. To control the overall one-sided type I error under 0.025, the cumulative alpha spending at the three analyses will be 0.0001, 0.0249, and 0.0250, respectively. As a result, the p-value stopping boundaries for the three analyses will be 0.0001, 0.0249 and 0.0003, respectively. The statistical significance of OS will be claimed as specified in Section 6.3.

4 ANALYSIS SETS

4.1 Intent-To-Treat (ITT) Analysis Set

The intent-to-treat (ITT) analysis set includes all subjects who are randomized, with treatment group designated according to randomization.

This analysis set will serve as the primary analysis set for all efficacy endpoints and demographics.

4.2 All-Treated (AT) Analysis Set

The all-treated (AT) analysis set includes all subjects who receive any amount of study drug (duvelisib or ofatumumab), with treatment group designated according to actual study treatment received. The AT analysis set will be the primary analysis set for safety endpoints.

4.3 Per-Protocol (PP) Analysis Set

The per-protocol (PP) analysis set includes all subjects in the ITT analysis set who do not violate the terms of the protocol in a way that would significantly affect the study outcome, with treatment group designated according to randomization. Subjects who meet any of the following criteria may be excluded from this analysis set:

- Do not have documented diagnosis of CLL or SLL
- Do not have measurable nodal disease at baseline as determined by the IRC
- ECOG performance status >2
- History of Richter's transformation or prolymphocytic leukemia
- Refractory to ofatumumab (defined as progression or relapse <12 months of receiving ofatumumab monotherapy or <24 months of receiving an ofatumumab-containing regimen)
- Prior exposure to a PI3K inhibitor
- Receive concomitant prohibited anticancer therapy
- Permanent discontinuation from study drug due to non-compliance

The PP analysis set will be a secondary analysis set for selected efficacy analyses.

5 STUDY ENDPOINTS

5.1 Primary Endpoint

The primary efficacy endpoint PFS is defined as time from randomization to the first documentation of progressive disease (PD) as determined by independent review or death due to any cause.

The censoring method of the primary endpoint can be found in Appendix A. The detailed algorithm will be specified in the Analysis Data Model (ADaM) specifications.

5.2 Secondary Endpoints

The secondary efficacy endpoints of the study are:

- Overall Response Rate (ORR), with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR), or PR with lymphocytosis (PRwL), according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Lymph node response rate, with lymph node response defined as ≥50% decrease in the SPD of target lymph nodes
- Overall Survival (OS), defined as time from randomization to death
- Hematologic improvement rate, defined as any of following hematologic improvement sustained for at least 60 days without transfusion or exogenous growth factors:
    - Neutrophil count >1,500/µL or an increase ≥50% from Baseline; or
    - Hemoglobin >11 g/dL or an increase ≥50% from Baseline; or
    - Platelet count >100,000/µL or an increase ≥50% from Baseline
- Duration of Response (DOR), defined as time from the first documentation of response to first documentation of PD or death due to any cause.

Other secondary endpoints of the study are:

- TEAEs and changes in safety laboratory values
- PK parameters derived from plasma duvelisib concentrations and, if applicable, its metabolite(s)

5.3 Exploratory Endpoints

The exploratory efficacy endpoints of the study are:

- Improvement in disease-related symptoms of fatigue, weight loss, unexplained fevers and night sweats from Baseline
- Minimal Residual Disease (MRD) in subjects with documented CR or CRi
- Time to Response (TTR), defined as the time from randomization to the first documentation of response (CR, CRi, PR or PRwl)

The QoL endpoints of the study are:

- Health-related QoL of subjects as assessed by the following subject-reported questionnaires:
    - EuroQol-5D (EQ-5D)
    - Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F)

The biomarkers of the study are:

- Serum and tissue biomarkers and blood immunophenotype
- Tumor genomic features (eg, DNA sequence variation, DNA copy number variation, and/or RNA expression)
- Germline DNA sequence variations

6 GENERAL STATISTICAL METHODS AND DATA HANDLING

6.1 General Methods

Summary statistics will be presented by treatment group, unless stated otherwise.

Unless otherwise specified, descriptive statistics for continuous data will include the number of subjects with data to be summarized (n), mean, standard deviation, 25% quartile, median, 75% quartile, and minimum and maximum. The same number of decimal places as in the raw data will be presented when reporting the minimum and maximum, one more decimal place than the raw data will be presented when reporting mean and median, and 2 more decimal places than the raw data will be presented when reporting standard deviation.

Descriptive statistics for categorical/qualitative data will include frequency counts and percentages. The total number of subjects in the treatment group will be used as the denominator for percent calculations, unless stated otherwise. All percentages will be presented with one decimal, unless otherwise specified. Percentages equal to 100 will be presented as 100, and percentages will not be presented for zero frequencies.

Descriptive statistics associated with time-to-event analyses will include the number of events, the number of subjects censored, 25% quartile, median, 75% quartile, and 95% confidence interval for median. These statistics will be presented for all time-to-event analyses, unless stated otherwise.

Listings will be provided for selected endpoints. For listings broken down by center and treatment arm, site (center) number will be ordered by country.

6.2 Handling of Missing Data

In general, values for missing data will not be imputed unless methods for handling missing data are specified.

6.2.1 Handling of Missing Dates/Months/Years for Adverse Events

Adverse events (AEs) with incomplete onset dates will be handled as follows for the sole purpose of determining treatment emergence (TEAE is defined in Section 7.3.1):

- If the start/end date of an AE is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The AE will be assumed to be treatment emergent if it cannot be definitively shown that the AE did not occur or worsen during the treatment-emergent period (worst case approach). The detailed algorithm will be specified in ADaM specifications.
- If the start date is completely missing, an AE will be considered treatment-emergent unless the stop date is before study drug administration.
- If the dose start date is missing for a subject at a data-cut, all AEs of the subject will be considered treatment-emergent.

The original partial or missing date will be shown in listings of AEs.

6.2.2 Handling of Missing Dates/Months/Years for Concomitant Medications

Prior or concomitant medications with incomplete start dates will be handled as follows for the sole purpose of determining whether a non-study medication is a concomitant medication:

- If the start/stop date of a medication is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The medication will be assumed to be concomitant if it cannot be definitively shown that the stop date is before the start of administration of study drug, or the start date is more than 30 days after the last date of administration of study drug. The detailed algorithm will be specified in ADaM specifications.
- If the start/stop dates are completely missing, a medication will be considered concomitant.
- If the dose start date is missing for a subject at a data-cut, all non-study medications of the subject will be considered concomitant.

The original partial or missing date will be shown in listings of all non-study medications.

6.2.3 Handling of Missing Dates/Months/Years for Disease History and Prior Therapies

For the purpose of calculating the duration from initial diagnosis, most recent relapse/refractory diagnosis or most recent prior therapy to randomization, partial/missing dates for diagnosis and last prior therapy completion will be imputed as follows:

- If both date and month are missing and the year is prior to the year of screening, the imputed date and month will be 01 July.
- If both date and month are missing and the year is the same as the year of screening, the imputed date will be the middle point between 01 Jan of the year and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- If date is missing and the month and year are prior to the month and year of screening, the imputed date will be 15th day of the month.
- If date is missing and the month and year are the same as the month and year of screening, the imputed date will be the middle point between the first date of the month and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- No imputation will be performed if the year is missing.

6.3 Multiple Comparisons/Multiplicity Adjustment

Of the secondary endpoints, ORR, Lymph Node Response Rate, OS, and Hematologic Improvement Rate are designated as key secondary efficacy endpoints. The primary endpoint and key secondary endpoints will be tested at an overall one-sided alpha level of 0.025 based on a gatekeeping approach.

If the primary endpoint is significant, the 4 secondary endpoints will be sequentially tested in the order listed above. ORR will be tested at the one-sided 0.025 level only if PFS is declared

statistically significant. Lymph Node Response Rate will be tested at the one-sided 0.025 level only if PFS and ORR are declared statistically significant. OS will be tested according to the planned alpha spending specified in Section 3.5.2 only if PFS, ORR and Lymph Node Response Rate are declared statistically significant. Hematologic Improvement Rate will be tested at the one-sided 0.025 level only if PFS, ORR, Lymph Node Response Rate and OS are declared statistically significant. If any null hypothesis is not rejected in this sequence of tests, formal sequential testing will be stopped.

6.4 Adjustments for Covariates

Adjustments for covariates will be considered for analysis of primary and key secondary endpoints, with details provided in Sections 7.2.1 and 7.2.2.

6.5 Subgroups

PFS, ORR and OS will be examined in the subgroups based on the following baseline characteristics or variables:

- Stratification factors (1. captured from central lab; 2 and 3 captured on the eCRF):
    1 High-risk cytogenetics (presence vs absence of del[17p])
    2 Refractory/early relapse to purine analog-based therapy (defined as progression <12 months after fludarabine/pentostatin: yes vs no)
    3 Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)
- Diagnosis (CLL or SLL)
- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Previously treated with ofatumumab (yes vs no, if sample size allows the analysis)
- Time from last dose of most recent prior anti-cancer therapy to randomization <12 months: yes vs no

If TP53 mutation status is available at the time of analyses, PFS, ORR and OS may also be examined in the subgroups based on the following variable:

- del[17p] or TP53 mutation (either or both present vs neither present; del[17p] will be as captured from central lab)

TEAEs (All Causalities) and TEAEs (Treatment-Related) will be examined in the following subgroups:

- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Baseline ECOG performance status (0 or 1 vs 2)

More details will be specified in Sections 7.2.1, 7.2.2 and 7.3.1.

6.6 Visit Windows

All data will be categorized based on the scheduled visit at which it was collected. These visit designators are predefined values that appear as part of the visit tab in the eCRF. There will be no additional analysis windowing done based on the assessment date.

6.7 Unscheduled Visits

Unscheduled visits will not be included in by-visit summary tables, unless otherwise specified. For laboratory tests, data from unscheduled visits will be included in listings and summaries of maximum changes from baseline, and the best or worst post-baseline values. For endpoints based on disease status assessment, data from unscheduled assessments will be included in the derivation and analyses of the endpoints.

6.8 Baseline Values

Unless otherwise specified, the baseline value is defined as the value collected at the time closest to, and prior to, the start of study drug administration. Values collected at unscheduled visits prior to the start of the study drug administration will be included in the calculation of baseline values.

6.9 Computing and Coding Standards

Activities will be performed using the following tools:

Table, listing, and figure production	SAS Version 9.2 or higher
Coding	
Adverse Events	MedDRA Version 16.1 or higher
Medical Histories	MedDRA Version 16.1 or higher
Prior and Concomitant Medications	WHODrug Version September 2013
Grading	
AEs	CTCAE Version 4.03
Labs	CTCAE Version 4.03

7 STATISTICAL ANALYSES

7.1 Study Subjects

7.1.1 Disposition of Subjects

The disposition of subjects will include the number and percentage of subjects for the following categories: the number randomized, the number and percentage randomized but not dosed, the number and percentage dosed. These categories will be summarized for each treatment arm and for the two treatment arms combined (total). The percentages will be based on all randomized subjects (ITT analysis set).

An end-of-treatment disposition (still on treatment vs discontinued from treatment) will be provided for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set. The primary reason for treatment discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

An end-of-study disposition (still on study vs discontinued from study) will also be provided for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The primary reason for study discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

A summary of strata as captured in IRT will be presented by treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. A summary table and a listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF will be presented.

7.1.2 Protocol Deviations

Protocol deviations will be categorized as major or minor prior to data release for the interim or final analysis of the primary endpoint. A summary table of the major protocol deviations will be provided by treatment arm and for two arms combined (total) for the ITT analysis set. A listing of major protocol deviations will be provided, broken down by center and treatment arm.

7.1.3 Demographic and Other Baseline Characteristics

Demographic and other baseline variables will be summarized for each treatment arm based on all randomized subjects (ITT analysis set) and for the two treatment arms combined (total). The variables will include age, age group (<65 versus >=65), sex, race, ethnicity, height, weight and the following biomarkers:

- del[17p] (captured from central lab, presence vs absence vs indeterminate)
- del[17p] (captured from local lab, presence vs absence vs indeterminate)
- IgHV status (mutated vs unmutated vs indeterminate)
- CD38 (Positive (≥30%) vs Negative (<30%) vs indeterminate)
- ZAP70 (Positive (>19%) vs Negative (≤19%) vs indeterminate)

A separate table will be provided to compare del[17p] values captured from central lab and local lab. Demographic and other baseline variables will also be summarized based on the PP analysis set.

TP53 mutation status (biomaker) may be summarized as a baseline characteristic as described above, if data are availble at the time of analyses. TP53 may be summarized alone and/or in association with del[17p]:

- del[17p] and/or TP53 mutation (either or both present vs neither present vs indeterminate; del[17p] will be as captured from central lab)
- TP53 mutation (presence vs absence vs indeterminate)

7.1.4 Disease History

Disease history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include diagnosis (CLL or SLL), risk factors used for randomization that are reported on eCRF (high risk cytogenetics [17p deletion: presence vs absence], refractory/early relapse to purine analog based treatment [yes vs no], grade 4 cytopenia(s) [presence vs absence]), years from initial diagnosis to randomization, months from most recent relapse/refractory diagnosis to randomization, Binet/Rai stage at initial diagnosis, type of prior treatment, Binet/Rai stage at baseline, and baseline lymphocyte count.

The durations to be summarized are defined as follows:

- Years from initial diagnosis to randomization will be calculated as (date of randomization – date of initial diagnosis + 1)/365.25.
- Months from most recent relapse/refractory diagnosis to randomization will be calculated as (date of randomization – date of most recent relapse/refractory diagnosis + 1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.5 Prior Therapies

Prior therapies will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include number of prior systemic therapies (summarized as a continuous variable and as a categorical variable) and months from most recent prior systemic therapy to randomization.

The duration to be summarized is defined as follows.

- Months from most recent prior systemic therapy to randomization will be calculated as (date of randomization – stop date of most recent systemic therapy +1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.6 Medical History

Medical history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set.

Medical history will be summarized by system organ class (SOC) and preferred term (PT) using the number and percentage of subjects who had at least one occurrence of an SOC or PT. The summary will be sorted alphabetically in SOC and by decreasing frequency of PT in the duvelisib arm within an SOC.

7.1.7 Prior and Concomitant Medications

Medications will be considered as prior if they stopped before the date of first dose of study drug.

Medications will be considered concomitant if they were taken at any time between the date of first dose of study drug and 30 days after the date of last dose of study drug, inclusive. If the start date or end date of a medication is completely or partially missing, refer to Section 6.2.2 for the algorithm to determine whether a medication is concomitant.

Prior medications and concomitant medications will be summarized separately. Both summaries will be based on the AT analysis set.

Medications will be summarized by ATC level 1, ATC level 2, and preferred drug name for each treatment arm. The summary will be sorted by decreasing frequency in ATC level 1, ATC level 2 and preferred drug name in the duvelisib arm. A subject taking the same drug multiple times will only be counted once.

A listing will be provided for all non-study medications taken on the study. An identifier will be provided to show if a medication is prior or concomitant. Medications that started more than 30 days after the last dose of study drug will be identified as post-treatment.

7.1.8 Exposure to Study Drug

Extent of exposure will be summarized for each treatment arm based on the AT analysis set.

Extent of exposure will be summarized for the following variables:

- Duration (weeks): (date of last dose – date of first dose + 1) divided by 7
- Number of cycles started (continuous and categorical)
- Ofatumumab arm only: For each infusion (1-12), number and percent of subjects receiving the infusion
- Relative dose intensity, defined as 100% x (total dose received)/ (planned cumulative dose for the duration of treatment)
- Number and percentage of subjects with a dose reduction
- duvelisib arm only: Number and percentage of subjects with a dose increase
- Number and percentage of subjects with a dose interruption

- Number and percentage of subjects with study drug discontinued

A by-subject listing will be presented for exposure to study drug.

7.2 Efficacy Analyses

For PFS, ORR, lymph node response rate, DOR and other endpoints derived from progression and/or response status, the primary analyses will be based on the endpoints derived from independent central review by Independent Review Committee (IRC). The endpoints derived from investigator assessment will be used in sensitivity analyses.

All efficacy analyses will be based on the ITT analysis set unless stated otherwise. If analyses are performed on more than one analysis set, the analyses on the ITT analysis set will be considered primary.

For stratified analysis of any efficacy endpoint, the following algorithm will be used to pool strata if there is insufficient information in any stratum (ie, there are <6 subjects, or there is no event for a time-to-event endpoint, or all subject have the same outcome for a binary endpoint in a stratum). (1) Strata will be ranked from the smallest to the largest based on the number of subjects, and if there is a tie, based on the number of events or responses. If there is still a tie, based on the reverse order of strata as determined by the stratification factors and levels (see Section 3.3). (2) The smallest stratum will be compared with the criteria of insufficient information. (3) If there is insufficient information in the smallest stratum, that stratum will be pooled with the smallest of the adjacent strata, which are defined as strata having 2 of the 3 stratification factors being at the same level as the smallest stratum. At the end of the three steps, the pooled stratum will replace the original two contributing strata. If there is still insufficient information in any stratum, the three steps will be repeated with the last pooled stratum assuming the stratum label of the larger of the two contributing strata for the purpose of additional pooling. The resulting pooled strata will be the strata used for the stratified analysis described in this section.

Listings of all efficacy endpoints will be provided.

7.2.1 Analyses of Primary Endpoint

7.2.1.1 Primary Analyses of PFS

The primary analyses of PFS will use PFS based on IRC assessment in the ITT analysis set.

Number of subjects with events, types of events (progression or death before progression), number of subjects censored, number of subjects for each reason of censoring, estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points may also be presented.

A stratified log-rank test (one-sided) will be used to compare PFS of the duvelisib arm against PFS of the ofatumumab arm at the interim and final analyses with the overall one-sided significance level controlled at 0.025. The nominal p-value required for statistical significance at

an analysis will be calculated based on the numbers of PFS events at all preceding and current analyses. The HR (duvelisib/ofatumumab) and the corresponding 2-sided 95% CI will be estimated using a stratified Cox proportional hazards model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

PFS will be plotted for each treatment group using the Kaplan-Meier method. The plot will include the number of subjects at risk for each treatment group over time.

7.2.1.2 Sensitivity Analyses of PFS

The following sensitivity analyses will be performed:

- PFS based on investigator assessment: This endpoint will be analyzed using the same methods as the primary analyses described for PFS based on IRC assessment. In addition, the differences between investigator assessment and IRC assessment will be summarized.
- Worst-case sensitivity analysis: Subjects who are alive and have not had documented progression by data cutoff and who are "lost to follow-up" (missing at least one disease assessment right before data cut off) will be treated as censored at their last adequate disease assessment if they are on the control arm (ofatumumab) and treated as having a PFS event at the time of the next scheduled assessment following the last adequate disease assessment if they are on the experimental arm (duvelisib). PFS based on IRC assessment with the above worst-case censoring/event rule will be analyzed using the same methods as the primary analyses for PFS.
- Event-free survival (EFS): This is defined as time from randomization to the first documentation of PD as determined by IRC, start of new anticancer treatment or procedure, or death due to any cause. The event/censoring method is presented in Appendix B. EFS will be analyzed using the same methods as the primary analyses for PFS except that the summary of types of events will include a category of new anticancer treatment or procedure.
- Analysis using the PP analysis set.
- Analysis using the AT analysis set.
- Unstratified analyses: An unstratified log-rank test (one-sided) will be used to compare the two trreatment arms. An unstratified Cox regression (one-sided) will be used to estimate the hazard ratio (duvelisib/ofatumumab) with its 95% confidence interval. Other aspects of the analyses will be the same as the primary analyses of PFS.
- Cox regression with baseline covariates: A stratified Cox regression will be used to test treatment effect on PFS, adjusting for demographic and other baseline characteristics. The strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. A stepwise variable selection will be performed to choose the variables in the Cox regression. Candidate variables are age, gender, race, disease diagnosis (CLL or SLL), years from initial

diagnosis, months from most recent relapse/refractory diagnosis, stage at diagnosis, stage at baseline, and number of prior systemic therapies. PFS based on IRC assessment will be used for this analysis.

7.2.1.3 Subgroup analyses of PFS

Subgroup analyses for PFS will be performed using the subgroups specified in Section 6.5. The HR (duvelisib/ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The hazard ratios and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

7.2.2 Analyses of Secondary Efficacy Endpoints

Of the secondary endpoints, ORR, Lymph Node Response Rate, OS, and Hematologic Improvement Rate are designated as key secondary efficacy endpoints. These endpoints will be tested for statistical significance under an overall one-sided significance level of 0.025 if and only if the primary endpoint is significant. Details of multiplicity adjustment are provided in Section 6.3.

The analysis of another secondary efficacy endpoint, Duration of Response (DOR), will be descriptive only. No hypothesis testing will be performed.

7.2.2.1 Overall Response Rate (ORR)

The primary analyses of ORR will use ORR based on IRC assessment in the ITT analysis set.

ORR will be derived from best overall response (BOR), which is defined as the best time point response that a subject achieves during the course of the study, with the response ranked according to the following order (from best to worst): CR>CRi>PR>PRwL>SD>PD (CRi applies to CLL only). In addition, the IRC may assign a timepoint response of unknown (UNK) due to missing, incomplete, or inadequate data; at baseline, the IRC may assign a timepoint response of no evidence of disease (NED) if both radiological and clinical data indicate no disease involvement; for a post-baseline timepoint, the IRC may also assign a timepoint response of not evaluable (NE) if no target lesions were identified at baseline and the radiological and clinical data at the post-baseline timepoint do not support the disease response of PD or UNK; these categories will be classified as OTHER in the summary.

The estimated ORR (percent of subjects with a BOR of CR, CRi, PR or PRwL) and a 2-sided 95% CI will be provided for each treatment arm. Number and percent of subjects with BOR in each of the response categories (CR, CRi, PR, PRwL, SD, PD, OTHER) will also be presented by treatment arm. All subjects in the analysis set will be included in the denominator in the calculation of the percentage for each response category or ORR.

ORR will be analyzed using the Cochran-Mantel-Haenszel (CMH) test (one-sided, see Appendix C) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. The odds ratio and its 95% CI will be estimated.

The following sensitivity analyses will be performed:

- ORR based on investigator assessment
- Overall Confirmed Response Rate (OCRR), with overall confirmed response (based on independent review) defined as best confirmed response (time between response and confirmation must be ≥8 weeks in duration) of CR, CRi, PR, or PRwL, according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Overall Response Rate without PRwL: with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR)
- Overall Confirmed Response Rate without PRwL
- Overall Response Rate in subset of subjects with baseline assessment other than UNK and NED. This analysis will be performed if there are at least 5% subjects with baseline assessment of UNK or NED
- Overall Response Rate in subset of subjects who are on treatment for 60 days or longer
- Analysis using the PP analysis set
- Analysis using the AT analysis set

Subgroup analyses will be performed for ORR using the subgroups specified in Section 6.5.

The odds ratio (duvelisib/ofatumumab) with its 95% CI will be calculated for each subgroup based on unstratified test. The odds ratios and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

7.2.2.2 Lymph Node Response Rate

The primary analyses of lymph node response rate will be based on IRC assessment in the ITT analysis set. The number of subjects with lymph node response, estimated lymph node response rate, and a 2-sided 95% CI will be provided for each treatment arm.

Lymph node response rate will be analyzed using the CMH test (one-sided) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

A sensitivity analysis will be performed using lymph node response rate based on investigator measurements.

7.2.2.3 Overall Survival (OS)

The primary analyses of OS will be based on the ITT analysis set.

Subjects without documentation of death at the time of the data cutoff for analysis will be censored at the date the subject was last known to be alive, or the data cutoff date, whichever is earlier. A stratified one-sided log-rank test will be used to compare OS between the 2 treatment groups. The HR along with the 95% CI will be estimated using a stratified Cox model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified

randomization, with potential pooling of strata specified earlier in Section 7.2. A Kaplan-Meier plot for OS will be presented by treatment group. Estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for OS will be presented by treatment group (if estimable). Probabilities of survival at selected time points may also be presented.

A majority of subjects can be expected to take subsequent anticancer therapy, especially after disease progression. To adjust for the effects of subsequent therapy, the following sensitivity analysis may be performed:

1. Primary analyses except additional censoring at the start date of subsequent therapy
2. Rank preserving structural failure time (RPSFT) analysis
3. Inverse probability of censoring weighted (IPCW) analysis

The following sensitivity analyses will be performed:

- Analysis using the PP analysis set.
- Analysis using the AT analysis set.
- Unstratified analyses: An unstratified log-rank test (one-sided) will be used to compare the two treatment arms. An unstratified Cox regression (one-sided) will be used to estimate the hazard ratio (duvelisib /ofatumumab) with its 95% confidence interval.

Subgroup analyses for OS will be performed using the subgroups specified in section 6.5. The HR (duvelisib /ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The HRs and their 95% CIs will be displayed for all subgroups graphically in a forest plot.

OS follow-up time, defined as time from randomization to the date of last procedure or assessment (including telephone interview), will be summarized by treatment arm and for the two treatment arms combined (total) using the reverse Kaplan-Meier method, which applies Kaplan-Meier method to the OS data with the same event/censoring time as the primary OS analyses while reversing event/censoring indicator. The 25th percentile, median, and 75th percentile ofOS follow-up time will be presented.

7.2.2.4 Hematologic Improvement Rate

A subject with a hematologic improvement is one who consistently met the criteria of an improvement in neutrophil count, hemoglobin or platelet count for a period of at least 60 days during which the subject did not have a transfusion or exogenous cytokines. Missing data or missing scheduled assessments during a potential 60-day hematologic improvement episode will result in categorization of not having a hematologic improvement.

The number of subjects with hematologic improvement, estimated hematologic improvement rate, and a 2-sided 95% CI will be provided for each treatment arm.

Hematologic improvement rate will be compared between treatment groups using one-sided CMH test. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

7.2.2.5 Duration of Response (DOR)

DOR will be presented using the Kaplan-Meier method for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL) as determined by IRC assessment. The censoring method will be the same as that for the primary endpoint (see Appendix A). The Kaplan-Meier plot for DOR will be presented by treatment group. Estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for DOR will be presented by treatment group (if estimable). No treatment comparison will be performed.

7.2.3 Analyses of Exploratory Efficacy Endpoints

7.2.3.1 Improvement in Disease-Related Symptoms

Disease-related symptoms of fever, weight loss, and drenching night sweats will be assessed as present or absent at each timepoint. Improvement in each of the three disease-related symptoms from baseline will be analyzed separately. The summary will display by treatment arm the proportion of subjects with an improvement while on treatment. Descriptive analysis of improvement in these disease-related symptoms will also be presented over the course of the study (eg, for each scheduled assessment) by treatment arm.

The disease-related symptom of fatigue will be measured by the ECOG performance status. Summary statistics will be provided by treatment arm for the ECOG score and the change from baseline at each timepoint.

7.2.3.2 Minimal Residual Disease

MRD status for subjects with a response of CR or CRi will be determined using flow cytometry from a central lab. MRD will be classified as positive or negative. The estimated rates of subjects with negative MRD and their 95% confidence intervals in subjects with a response of CR or CRi will be provided for each treatment arm.

7.2.3.3 Time To Response (TTR)

TTR will be presented for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL). Summary statistics and a 2-sided confidence interval for median TTR will be provided by treatment arm. No treatment comparison will be performed.

7.3 Safety Analyses

All safety analyses will be performed using the AT analysis set.

7.3.1 Adverse Events

Adverse events will be coded using MedDRA Version 16.1 or higher. The Grade of the AE will be assessed according to the NCI-CTCAE Version 4.03. If an AE is not included in the NCI-CTCAE Version 4.03, the Grade of the AE will be assessed according to the protocol, Section ***8.2.1.2***.

The summary of AEs will be focused on treatment-emergent AEs. A treatment-emergent AE (TEAE) is defined as any AE that emerges or worsens in the period from the first dose of study treatment to 30 days after the last dose of study treatment. The onset date of an AE will be compared to the first dose date and the last dose date plus 30 days to determine whether the AE is treatment-emergent or not. If the onset date of an AE is completely or partially missing, refer to Section 6.2.1 for the algorithm to determine whether an AE is treatment emergent.

TEAEs will be summarized for each treatment arm by MedDRA system organ class (SOC) and preferred term (PT), or PT only. For summary tables by SOC and PT, SOC will be sorted alphabetically and PT will be sorted by decreasing frequency in the duvelisib arm within each SOC. For summary tables by PT only, PT will be sorted by decreasing frequency in the duvelisib arm.

If multiple TEAEs of the same PT occur within a subject, only the maximum grade observed for this PT will be used in summary of TEAEs by grade, the subject will be counted only once in the number of subjects for this PT and only once for the number of subjects for the SOC to which this PT belongs.

An overview TEAE summary table will be provided, which will include the number of subjects with AEs in selected categories. In addition, TEAEs will be summarized for the following categories, and will be tabulated by SOC and PT, unless otherwise specified.

- Treatment-emergent AEs (All Causalities)
- Treatment-emergent AEs (Treatment-Related)
- Treatment-emergent AEs (All Causalities, by maximum grade)
- Treatment-emergent AEs (Treatment-Related, by maximum grade)
- Grade 3 or higher treatment-emergent AEs (All Causalities)
- Grade 3 or higher treatment-emergent AEs (Treatment-Related)
- Treatment-emergent SAE (All Causalities)
- Treatment-emergent SAE (Treatment-Related)
- Treatment-emergent AEs resulting in discontinuation of study drug (All Causalities)
- Treatment-emergent AEs resulting in dose hold (All Causalities)
- Treatment-emergent AEs resulting in dose reduction (All Causalities)
- Treatment-emergent AEs resulting in dose hold or reduction (All Causalities)

- Treatment-emergent AEs resulting in discontinuation of study drug (Treatment-Related)
- Treatment-emergent AEs resulting in death (All Causalities)
- Treatment-emergent AEs resulting in death (Treatment-Related)
- Treatment-emergent AEs by PT (All Causalities)
- Treatment-emergent AEs by PT (AEs with onset date within 24 weeks after the first dose of study treatment; All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (AEs with onset date within 24 weeks after the first dose of study treatment; All Causalities)
- Treatment-emergent SAEs by PT (All Causalities)
- Treatment-emergent SAEs by PT (Treatment-Related)
- Treatment-emergent AEs reported in ≥5% of subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)

A by-subject listing of the following AE categories will be presented.

- All AEs (TEAEs will be flagged)
- All SAEs (TEAEs will be flagged)
- Treatment-emergent AEs resulting in dose hold
- Treatment-emergent AEs resulting in dose reduction
- Treatment-emergent AEs resulting in discontinuation of study drug
- Treatment-emergent AEs resulting in death

Treatment-emergent AEs (All Causalities) and Treatment-emergent AEs (Treatment-Related) will also be tabulated in each subgroup specified in Section 6.5.

7.3.2 Laboratory Data

Laboratory tests will be reported separately for hematology and blood chemistry.

For the purposes of presentation in both tables and listings, the following laboratory test results will be converted to the International System of Units (SI) before presentation: sodium, potassium, chloride, bicarbonate (or CO_2), albumin, total protein, creatinine, uric acid, calcium, phosphorus, magnesium, glucose, total and direct bilirubin, and alkaline phosphatase, red blood cell (RBC) count, hemoglobin, hematocrit, platelets, white blood cell count with 5-part differential performed manually or by flow cytometry (lymphocytes, neutrophils, monocytes, basophils, and eosinophils), etc.

Lab tests performed after the start of the first dose of study treatment, up to 30 days from the last dose, will be considered treatment-emergent. Only treatment-emegent lab tests will be included in the analyses in this section.

If a laboratory test value is reported using a non-numeric qualifier (e.g., less than [<] a certain value, or greater than [>] a certain value), the given numeric value will be used in the summary statistics, ignoring the non-numeric qualifier.

For laboratory tests with NCI-CTCAE grades, a shift table from baseline grade to the maximum post-baseline grade will be provided. Laboratory tests with bi-directional grades (e.g., Hyperglycemia and Hypoglycemia) will be presented separately for each direction within the shift table.

Listings will be provided for all laboratory test results and for laboratory test results grade 3 and higher. A listing of subjects with ALT or AST >3xULN with simultaneous total bilirubin >2xULN will be presented, where ULN stands for upper limit of normal. The elevations of ALT/AST and total bilirubin must occur within 2 days of each other.

7.3.3 Vital Signs

The actual values of vital sign parameters, including temperature, heart rate, weight and systolic and diastolic blood pressure, will be presented in a by-subject listing.

7.3.4 Electrocardiogram (ECG)

A by-subject listing will be presented for baseline ECGs and post-baseline unscheduled ECGs (if any).

7.3.5 Concomitant Medications and Procedures

Please refer to Section 7.1.7 for the definition and summary of concomitant medications.

Concomitant procedures will not be summarized. A by-subject listing will be presented.

7.3.6 Deaths

On-treatment deaths are defined as deaths that occur within 30 days of the last dose of duvelisib. Deaths on treatment and causes for deaths occurring on-treatment, and deaths during follow-up will be summarized.

A by-subject listing of deaths on study will be presented.

7.4 Quality of Life Instruments

Analyses of the QoL instruments (EQ-5D and FACIT-F) will be performed on the ITT analysis set. Additional analyses may be considered as needed.

7.4.1 EQ-5D

The EQ-5D contains a descriptive system with one response for each of 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). The first response is coded as 1 (indicating no problems), the second response is coded as a 2 (indicating some problems), and the third response is coded as a 3 (indicating extreme problems). Ambiguous responses (eg, more than one response in a dimension) are treated as missing values. The EQ-5D also contains a Visual Analog Scale (VAS) for health state ranging from 0 to 100, where 0 represents worst imaginable health state and 100 represents the best imaginable health state.

For each dimension of the descriptive system, the number and percentage of subjects with no problems, some problems, and extreme problems will be reported at each visit.

Summary statistics for the VAS of health state and the change from baseline will be reported for each visit.

7.4.2 FACIT-F

The FACIT-F contains 5 subscales, Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, Functional Well-Being, and Fatigue (Additional Concerns on eCRF). Each subscale contains multiple items, and each item has a response of 0 to 4. Scores will be computed for each subscale, as well as for the FACIT-F Trial Outcome Index, FACT-G total score, and the FACIT-F total score.

The Physical Well-Being subscale contains 7 items. To compute the subscale score, compute the item score by subtracting each item response from 4 (item score = 4 – item response), sum the item scores, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Physical Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Social/Family Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Social/Family Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Emotional Well-Being subscale contains 6 items. To compute the subscale score, compute the item score first. For items 1, 3, 4, 5, and 6, subtract each item response from 4 (item score = 4 – item response). For item 2, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 6 and divide by the number of items answered to compute the subscale score. At least 3 questions must be answered. If fewer than 3 questions are answered, then the subscale score will be missing. The Emotional Well-Being score ranges from 0 to 24 with higher scores representing better quality of life.

The Functional Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will

be missing. The Functional Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Fatigue subscale contains 13 items. To compute the subscale score, compute the item score first. For all items except An5 and An7, subtract each item response from 4 (item score = 4 – item response). For items An5 and An7, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 13 and divide by the number of items answered to compute the subscale score. At least 7 questions must be answered. If fewer than 7 questions are answered, then the subscale score will be missing. The Fatigue score ranges from 0 to 52 with higher scores representing better quality of life.

To derive the FACIT-F Trial Outcome Index, sum the subscale scores for Physical Well-Being, Functional Well-Being, and Fatigue. If any of these subscale scores is missing, the score for the FACIT-F Trial Outcome Index will be missing. The score ranges from 0 to 108.

To derive the FACT-G score, sum the subscale scores for Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, and Functional Well-Being. If any of these subscale scores is missing, the FACT-G score will be missing. The score ranges from 0 to 108.

To derive the FACIT-F total score, sum all 5 subscale scores. If any subscale score is missing, the FACIT-F total score will be missing. The score ranges from 0 to 160.

For each of the five subscale scores and for each of the FACIT-F Trial Outcome Index, the FACT-G score, and the FACIT-F total score, summary statistics for the actual values and changes from Baseline will be reported at each visit.

7.5 Pharmacokinetic Analyses

Plasma samples will be analyzed for duvelisib and IPI-656 (metabolite) concentrations using a validated high performance liquid chromatography-tandem mass spectrometry (LC/MS/MS) method. A listing of the duvelisib and IPI-656 concentrations will be provided; the sample collection date and time, and the date, time and amount of the preceding duvelisib dose administration will be included in the listing.

The PK data collected will be analyzed by standard population PK methods, using appropriate software. Analysis of exposure-response relationships for efficacy and safety endpoints will be conducted. If there is only a limited amount of plasma concentration data from this study, the data may be pooled with the results of other studies to perform the population PK and exposure-response analyses. Further details on these analyses will be outlined in a separate analysis plan. Results of the population PK and exposure-response analysis will be summarized in a separate technical report.

7.6 Biomarker Analyses

The following exploratory analyses may or may not be presented formally. If presented, they will be limited to biomarkers for which data from a sufficient number of subjects are available.

Further details on these analyses may be outlined in a separate analysis plan, and results may be summarized in a separate techinical report.

7.6.1 Pharmacodynamic and Potentially Predictive Biomakers

The relationship between tumor genomics, gene expression, protein expression, and clinical endpoints may be explored as follows:

- Evaluation of serum biomarkers,blood immunophenotype and tumor biomarkers for relationships with duvelisib clinical activity and safety
- Evaluation of tumor genomic features (eg, cytogenetics, FISH, DNA sequence variation, DNA copy number variation, and/or RNA expression) for relationships with duvelisib clinical activity
- Evaluation of the relationship between serum and/or tumor biomarkers and disease progression of duvelisib and ofatumumab

7.6.2 Pharmacogenomics

DNA may be extracted from the pharmacogenomics sample in order to evaluate a subject's germline DNA. This sample may be used to explore the following:

- Evaluate the relationship between germline DNA sequence variations and PK of duvelisib
- Evaluate germline DNA as a control to verify that tumor sequence variations are somatic mutations

8 CHANGES IN PLANNED ANALYSES

8.1 Changes in Planned Analyses from Protocol Amendment 1 to SAP V1.0

There are no changes from Protocol Amendment 1. More details are provided in the SAP.

8.2 Changes in Planned Analyses from SAP V1.0 to SAP V2.0

The main changes from SAP V1.0 to SAP V2.0 are described below.

1. Section 3.5.2 Clarification

   Old (V1.0): The second analysis of OS will be performed at the PFS final analysis, which is expected to happen after 185 PFS events have occurred, or approximately 27.5 months after the first subject is randomized.

   New (V2.0): The second interim analysis of OS will be performed at the PFS final analysis, which is to happen after 185 PFS events have occurred. *If the planned final analysis of PFS is not performed due to early stop for efficacy, then the second interim analysis of OS will be conducted approximately one year after the first interim analysis of OS.*

2. New censoring method for event-free survival were specified.
3. Prior irradiation and prior surgery were removed from prior therapy summary.

The following additional analyses/summaries have been added to this version of the SAP:

1. Analysis of ORR of subjects who are on treatment for 60 days or longer
2. Median time to OS follow up
3. Time to response
4. Treatment-emergent AEs reported in ≥5% of CLL/SLL subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
5. Grade 3 or higher treatment-emergent AEs reported in ≥5% of CLL/SLL subjects and occurred at ≥2% higher incidence in subjects receiving duvelisib (All Causalities)
6. Treatment-emergent AEs by PT (All Causalities, AE onset date within 24 weeks after the first dose of study treatment)
7. Treatment-emergent AEs resulting in discontinuation of study drug (Treatment-Related)
8. Treatment-emergent AEs resulting in death (Treatment-Related)
9. Treatment-emergent SAEs by PT (All Causalities).
10. Treatment-emergent SAEs by PT (Treatment-Related)
11. Deaths on treatment
12. Deaths during follow-up
13. A by-subject listing of treatment-emergent AEs resulting in dose hold

14. A by-subject listing of treatment-emergent AEs resulting in dose reduction
15. A by-subject listing of deaths on study
16. A summary table and a listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF
17. Comparision of del[17p] values captured from central lab and local lab.
18. The following biomarkers will be summarized as baseline characteristics:
    - del[17p] (captured from central lab, presence vs absence vs indeterminate)
    - del[17p] (captured from local lab, presence vs absence vs indeterminate)
    - del[17p] and/or TP53 mutation (del[17p] and/or TP53 mutation, del[17p] captured from central lab, either or both present vs neither present vs indeterminate)
    - del[17p] and TP53 mutation (del[17p] and TP53 mutation,del[17p] captured from central lab, both present vs either vs neither present vs indeterminate)
    - TP53 mutation (presence vs absence vs indeterminate)
    - IgHV status (mutated vs unmutated vs indeterminate)
    - CD38 (Positive (≥30%) vs Negative vs indeterminate)
    - ZAP70 (Positive (>19%) vs Negative (<=19%) vs indeterminate)
19. Subgroup analyses of PFS, OS and ORR based on the following factors:
    - Time from last dose of most recent prior anti-cancer therapy to randomization <12 months: yes vs no
    - del[17p] (captured from central lab) or TP53 mutation (either or both present vs neither present)

9 REFERENCES

1 O'Brien PC, Fleming TR. *A multiple testing procedure for clinical trials.* Biometrics 1979; 35:549-556.

2 Robins JM. *Information recovery and bias adjustment in proportional hazard regression analysis of randomized trials using surrogate markers*. Proceedings of the Biopharmaceutical Section, American Statistical Association 1993; pp. 24-33.

3 Robins JM, Finkelstein D. *Correcting for Non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests*. Biometrics, 2000; 56(3):779-788.

4 Robins JM, Tsiatis, AA. *Correcting for noncompliance in randomized trials using rank preserving structural failure time models.* Communications in Statistics, 1991; 20, 2609-2631.

10 APPENDICES

10.1 Appendix A: Primary PFS Event/Censoring Method

Censoring of PFS will be performed as detailed in the table below.

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression or death before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death	Date of death	Event
Documented progression or death following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment or procedure started before documented progression	Date of last adequate disease status assessment	Censored

Note: Disease status assessment includes CT scans (chest, abdomen and pelvis), bone marrow aspirate and/or biopsy (may not be required of all subjects at all scheduled disease status assessments), CBC and differential count, focused physical examination, disease related constitutional symptoms for disease assessment, and ECOG performance status. An adequate baseline disease status assessment is any baseline disease status assessment that include WBC counts and CT scans. An adequate post-baseline disease status assessment is any disease status assessment for which a disease status (eg, CR, CRi, PR, PRwL, SD, and PD) is arrived per protocol-defined criteria by the IRC (for IRC assessment) or investigator (for investigator assessment).

10.2 Appendix B: Event-Free Survival (EFS) Event/Censoring Method

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death or new anticancer treatment/procedure occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression, death or new anticancer treatment/procedure started before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression, and no new anticancer treatment/ procedure started before documented progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death, and no new anticancer treatment/procedure started before death	Date of death	Event
Documented progression, death or new anticancer treatment/procedure following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment/procedure started before documented progression with ≤1 missing scheduled disease status assessment prior to new anticancer treatment/procedure	Start date of new anticancer treatment/procedure	Event

10.3 Appendix C: One-Sided Cochran–Mantel–Haenszel (CMH) Test

The Cochran–Mantel–Haenszel (CMH) test compares binary responses of two treatment groups, adjusting for stratification factors. In the CMH test, the data are arranged in a series of associated 2 × 2 contingency tables, the null hypothesis is that the observed response is independent of the treatment used in any 2 × 2 contingency table.

Let O_{hij} be the observed frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h and E_{hij} be the expected frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h, where

$$E_{hij} = \frac{(O_{hi1} + O_{hi2}) \cdot (O_{h1j} + O_{h2j})}{O_{h11} + O_{h12} + O_{h21} + O_{h22}}$$

Also, let P_{hij} be the probability of outcome being j given treatment being i (i=1 or 2, j= 1 or 2) in stratum h. To test the following hypothesis,

$$H_0 : P_{h11} = P_{h21} \text{ for all } h \in \{1, \ldots, H\}$$
versus
$$H_1 : P_{h11} > P_{h21} \text{ for at least one } h \in \{1, \ldots, H\}$$

The one-sided Cochran-Mantel-Haenszel Test Statistics is constructed as:

$$Z_{CMH} = \frac{\sum_{h=1}^{H} (O_{h11} - E_{h11})}{\sqrt{V_{11}}} \sim Z$$

Where

$$V_{11} = Var(\sum_{h=1}^{H} (O_{h11} - E_{h11})) = \sum_{h=1}^{H} \frac{E_{h11} \bullet E_{h22}}{O_{h11} + O_{h12} + O_{h21} + O_{h22} - 1}$$

The test statistic Z_{CMH} will be compared with standard normal distribution to obtain p-value of the CMH test.

STATISTICAL ANALYSIS PLAN

Protocol IPI-145-07

A Phase 3 Study of IPI-145 versus Ofatumumab in Subjects with Relapsed or Refractory Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma

Protocol Version:	**Amendment 1**
Type of Analysis Plan:	**Interim and Final Analysis**
Version:	**1.0**
Date:	**2 Oct 2014**
Author:	**Shijie Tang, PhD**

The undersigned has developed this statistical analysis plan (SAP):

Name/Title	Signature	Date
Shijie Tang, PhD Associate Director Biostatistics		02 Oct 2014

The undersigned have reviewed this SAP and approve it in its entirety:

Name/Title	Signature	Date
Patrick Kelly, MD Vice President Clinical Development		02 Oct 2014
Katherine Reid, MPH Manager Pharmacovigilance		02 Oct 2014
Justin, McLaughlin, MA Director Medical Writing		02 Oct 2014
Lixin Han, PhD Director Biostatistics		02 Oct 2014

DOCUMENT HISTORY

Version	Date	Author(s)	Brief Summary of Changes
1.0	2 Oct 2014	Shijie Tang	Original

Table of Contents

LIST OF ABBREIVATIONS

Abbreviation	Description
ADaM	Analysis Data Model
AE	Adverse Event
ANCOVA	Analysis of Covariance
AT	All-Treated
ALT	Alanine Aminotransferase
ANC	Absolute Neutrophil Count
AST	Aspartate Aminotransferase
ATC	Anatomical Therapeutic Chemical
BID	Twice a day
BOR	Best Overall Response
BUN	Blood Urea Nitrogen
CI	Confidence Interval
CLL	Chronic Lymphocytic Leukemia
CMH	Cochran–Mantel–Haenszel
CO2	Bicarbonate
CR	Complete Response
CRi	Complete Response with Incomplete Marrow Recovery
CS	Abnormal and Clinically Significant
CTCAE	Common Terminology Criteria for Adverse Events
DOR	Duration of Response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form
eDISH	Evaluation of Drug Induced Serious Hepatotoxicity
EQ-5D	European Quality of Life-5 Dimensions
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue
FACT-G	The Functional Assessment of Cancer Therapy - General

Abbreviation	Description
HR	Hazard Ratio
IDMC	Independent Data Monitoring Committee
IPI-145	(S)-3-(1-(9H-purin-6-ylamino)ethyl)-8-chloro-2-phenylisoquinolin-1(2H)-one
IPCW	Inverse Probability of Censoring Weighted
IRC	Independent Central Review
IRT	Interactive Response Technology
ITT	Intent-To-Treat
IV	Intravenous
IWCLL	International Workshop on Chronic Lymphocytic Leukemia
IWG	International Working Group
LC/MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LDH	Lactate Dehydrogenase
LFT	Liver Function Test
MedDRA	Medical Dictionary for Regulatory Activities
MEOI	Medical Events of interest
MMRM	Mixed Effect Repeated Measures
MRD	Minimal Residual Disease
NCI	National Cancer Institute
NCS	Abnormal but Not Clinically Significant
NED	No Evidence of Disease
ORR	Overall Response Rate
OS	Overall Survival
PD	Progressive Disease
PFS	Progression-Free Survival
PK	Pharmacokinetics
PP	Per-Protocol
PR	Partial Response

Abbreviation	Description
PRwL	PR with Lymphocytosis
PT	Preferred Term
QoL	Quality of Life
RBC	Red Blood Cell
RPSFT	Rank Preserving Structural Failure Time
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
SI	Standard International System of Units
SLL	Small Lymphocytic Lymphoma
SOC	System Organ Class
SPD	Sum of Products
TEAE	Treatment Emergent Adverse Event
UNK	Unknown
VAS	Visual Analog Scale
WHO	World Health Organization
WHO-DD	World Health Organization Drug Dictionary

1 INTRODUCTION

This is the statistical analysis plan (SAP) for study IPI-145-07, *A phase 3 study of IPI-145 versus ofatumumab in subjects with relapsed or refractory chronic lymphocytic leukemia/small lymphocytic lymphoma*. This SAP is prepared according to Amendment 1 of the protocol, dated 02 April 2014.

2 STUDY OBJECTIVES

2.1 Primary Objective

The primary objective of this trial is to examine the efficacy of IPI-145 monotherapy versus ofatumumab monotherapy in subjects with relapsed or refractory Chronic Lymphocytic (CLL) or Small Lymphocytic Lymphoma (SLL).

2.2 Secondary Objectives

- To determine the safety of IPI-145 in subjects with CLL or SLL
- To evaluate the pharmacokinetics (PK) of IPI-145 and, if applicable, its metabolite(s)

2.3 Exploratory Objectives

- To evaluate the health-related quality of life (QoL) of subjects
- To evaluate pharmacodynamic biomarkers of IPI-145
- To evaluate biomarkers that may predict IPI-145 clinical activity and/or safety
- To evaluate mechanisms of resistance in subjects who exhibit disease progression while being treated with IPI-145 or ofatumumab
- To evaluate genomic features of tumors predictive of response in subjects treated with IPI-145 or ofatumumab

3 STUDY DESIGN

3.1 Overview

Study IPI-145-07 is a two-arm, randomized (1:1 ratio), open-label, phase 3, superiority trial designed to evaluate the efficacy and safety of IPI-145 compared to ofatumumab.

Subjects who meet all the eligibility criteria at Screening will return to the clinic on Day 1 to receive their first dose of study drug (randomized to either IPI-145 or ofatumumab). The first treatment cycle for each treatment arm will be 3 weeks (21±2 days). Subsequent treatment cycles will be 4 weeks (28±4 days).

Subjects randomized to IPI-145 will be given a starting dose of 25 mg IPI-145 administered orally twice daily (BID) initially in a 21-day treatment cycle followed by 28-day treatment cycles for up to 18 cycles or until disease progression or unacceptable toxicity (whichever comes first). After 18 complete cycles of treatment, subjects may receive additional cycles of IPI-145 for up to 3 years (36 cycles of total treatment) if they have documented evidence of response and disease requiring continued treatment according to the modified IWCLL/revised IWG criteria.

Subjects randomized to ofatumumab will receive treatment consistent with the approved product labeling: 8 weekly infusions, starting with an initial IV dose of 300 mg ofatumumab on Day 1 followed by seven weekly doses of 2000 mg. Thereafter, subjects will receive 2000 mg ofatumumab once every month for four months or until disease progression or unacceptable toxicity (whichever comes first), as outlined in the schedule of assessments. Administration of ofatumumab will not exceed the 12 doses (within 7 cycles) as described in the prescribing information.

Subjects will be followed for survival for up to 3 years (or other duration as specified by the protocol) from randomization or until death. Follow-up visits will occur every 6 months which can be conducted through a telephone interview.

3.2 Sample Size Consideration

This study employs a randomized, open-label, parallel design to assess the potential superiority of IPI-145 treatment over ofatumumab treatment on PFS in CLL or SLL subjects.

The sample size was determined to provide sufficient power to test the primary endpoint of PFS in a group sequential design with one planned interim analysis.

Assuming an exponential distribution for PFS, a total of 185 PFS events (deaths or progressions) will provide approximately 93% power to detect a hazard ratio of 0.6 (median PFS of 9 months in the ofatumumab group versus 15 months in the IPI-145 group) using a 1-sided log-rank test at a 2.5% overall significance level (ie, alpha of 0.025). The study design employs the Lan-DeMets spending function for O'Brien-Fleming boundary as the alpha spending function and the Hwang-Shih-DeCani gamma (-4) spending function as the beta spending function. A total of 300 subjects will be randomized in a 1:1 ratio to receive either ofatumumab or IPI-145. Assuming a peak accrual rate of 30 subjects per month with 11 months for accrual ramp-up and a 4%

cumulative dropout rate per year, the enrollment would complete in 16 months, with the final analysis for PFS estimated to occur approximately 27.5 months from the randomization of the first subject.

3.3 Randomization

Once a subject has met all entry criteria, the Interactive Response Technology (IRT) will be used to generate a distinct subject identifier. If a subject discontinues from the study, the subject identifier will not be reused and the subject will not be allowed to re-enter the study.

Eligible subjects will be randomized via IRT in a 1:1 ratio to one of two treatment arms:

- Arm 1: IPI-145 25 mg BID
- Arm 2: Ofatumumab

In order to ensure subject balance between treatment groups, study subjects will be stratified by the following:

- High risk cytogenetics (presence vs absence of del[17p])
- Refractory/early relapse to purine analog based treatment (progression <12 months after fludarabine/pentostatin: yes vs no)
- Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)

Randomization is to occur within 7 days of first dose after all screening assessments have been completed.

3.4 Blinding

This study is open-label and Investigators, site staff, Sponsor and Sponsor designees will have access to treatment assignments for individual subjects. To reduce bias, a blinded independent central review of disease status will be conducted, which will be blinded to individual subject treatment assignments. Investigators, site staff, Sponsor and Sponsor designees will only have access to blinded (pooled) aggregate study data, except for a select number of Sponsor staff, who may review aggregate SAEs and Medical Events of Interest (MEOI) data by treatment groups.

3.5 Planned Analyses

For the interim analysis, the actual p-value boundaries will be calculated based on the actual number of PFS and OS events at the analysis. For the final analysis, the actual p-value boundary for efficacy will be calculated based on the actual number of PFS and OS events at both the interim analysis and the final analysis.

The Independent Data Monitoring Committee (IDMC) will review PFS, OS and other efficacy data at the time of PFS interim analysis. The list of IDMC deliverables, including tables and listings, is included in the IDMC charter. These deliverables will be based on the methods described in this SAP.

3.5.1 Planned Analyses for PFS

After approximately 50% of the planned PFS events (ie, 93 PFS events) have been observed based on the blinded, independent, central review, an interim analysis of efficacy will be performed.

At the interim analysis, PFS will be tested at the alpha level based on the Lan-DeMets alpha spending function for the O'Brien-Fleming boundary with the opportunity to stop the study (other than survival follow-up) for overwhelming evidence of efficacy. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for efficacy is 1-sided p-value of 0.0015 (corresponding approximately to a HR of 0.540). In the meantime, PFS will also be tested at the alpha level for futility based on the alpha and beta spending functions specified for the study design. If the interim analysis occurs at exactly 50% of the planned events, the criterion of stopping for futility is 1-sided p-value of 0.4791 (corresponding approximately to a HR of 0.990). The futility boundary of this study is non-binding, meaning that the Type I error is properly controlled even if the study is continued after the futility boundary for PFS is crossed at the interim analysis. Under the protocol design assumptions (ie, the study is fully enrolled in 16 months), the interim analysis is expected to occur approximately 17 months after the first subject is randomized.

If the study is not stopped at the interim analysis, the final analysis will be performed when approximately 185 PFS events have occurred in the study. The criterion of statistical significance (ie, boundary for efficacy) is 1-sided p-value of 0.0245 (corresponding approximately to a HR of 0.748).

3.5.2 Planned Analyses for OS

There will be three planned OS analyses. The first one will be at the planned PFS interim analysis, which is expected to happen after 93 PFS events have occurred, or approximately 17 months after the first subject is randomized. The second analysis of OS will be performed at the PFS final analysis, which is expected to happen after 185 PFS events have occurred, or approximately 27.5 months after the first subject is randomized. Since all subjects will be followed for survival for 6 years, the final analysis of OS will take place at the end of follow-up of all subjects, which is approximately 88 months after the first subject is randomized, or 6 years from last subject is randomized. To control the overall 1-sided type I error under 0.025, the cumulative alpha spending at the three analyses will be 0.0001, 0.0249, and 0.0250, respectively. As a result, the p-value stopping boundaries for the three analyses will be 0.0001, 0.0249 and 0.0003, respectively. The statistical significance of OS will not be claimed unless the PFS reaches statistical significance.

4 ANALYSIS SETS

4.1 Intent-To-Treat (ITT) Analysis Set

The intent-to-treat (ITT) analysis set includes all subjects who are randomized, with treatment group designated according to randomization.

This analysis set will be used in the analyses of subject characteristics and efficacy endpoints.

4.2 All-Treated (AT) Analysis Set

The all-treated (AT) analysis set includes all subjects who receive any amount of study drug (IPI-145 or ofatumumab), with treatment group designated according to actual study treatment received. The AT analysis set will be the primary analysis set for safety endpoints.

4.3 Per-Protocol (PP) Analysis Set

The per-protocol (PP) analysis set includes all subjects in the ITT analysis set who do not violate the terms of the protocol in a way that would significantly affect the study outcome, with treatment group designated according to randomization. Subjects who meet any of the following criteria may be excluded from this analysis set:

- Do not have documented diagnosis of CLL or SLL within medical records
- Do not have measurable nodal disease at baseline as determined by the IRC
- ECOG performance status >2
- History of Richter's transformation or prolymphocytic leukemia
- Refractory to ofatumumab (defined as progression or relapse <12 months of receiving ofatumumab monotherapy or <24 months of receiving an ofatumumab-containing regimen)
- Prior exposure to a PI3K inhibitor or a BTK inhibitor
- Receive concomitant prohibited anticancer therapy
- Permanent discontinuation from study drug due to non-compliance

The PP analysis set will be a secondary analysis set for selected efficacy analyses.

5 STUDY ENDPOINTS

5.1 Primary Endpoint

The primary efficacy endpoint PFS is defined as time from randomization to the first documentation of progressive disease (PD) as determined by independent review or death due to any cause.

The censoring rule of the primary endpoint can be found in Appendix A. The detailed algorithm will be specified in the Analysis Data Model (ADaM) specifications.

5.2 Secondary Endpoints

The secondary efficacy endpoints of the study are:

- Overall Response Rate (ORR), with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR), or PR with lymphocytosis (PRwL), according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Lymph node response rate, with lymph node response defined as ≥50% decrease in the SPD of target lymph nodes
- Overall Survival (OS), defined as time from randomization to death
- Hematologic improvement rate, defined as any of following hematologic improvement sustained for at least 60 days without transfusion or exogenous growth factors:
    - Neutrophil count >1,500/µL or an increase ≥50% from Baseline; or
    - Hemoglobin >11 g/dL or an increase ≥50% from Baseline; or
    - Platelet count >100,000/µL or an increase ≥50% from Baseline
- Duration of Response (DOR), defined as time from the first documentation of response to first documentation of PD or death due to any cause.

Other secondary endpoints of the study are:

- TEAEs and changes in safety laboratory values
- PK parameters derived from plasma IPI-145 concentrations and, if applicable, its metabolite(s)

5.3 Exploratory Endpoints

The exploratory efficacy endpoints of the study are:

- Improvement in disease-related symptoms of fatigue, weight loss, unexplained fevers and night sweats from Baseline
- Minimal Residual Disease (MRD) in subjects with documented CR or CRi

The QoL endpoints of the study are:

- Health-related QoL of subjects as assessed by the following subject-reported questionnaires:
    - EuroQol-5D (EQ-5D)
    - Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F)

The biomarkers of the study are:

- Serum and tissue biomarkers and blood immunophenotype
- Tumor genomic features (eg, DNA sequence variation, DNA copy number variation, and/or RNA expression)
- Germline DNA sequence variations

6 GENERAL STATISTICAL METHODS AND DATA HANDLING

6.1 General Methods

Summary statistics will be presented by treatment group, unless stated otherwise.

Unless otherwise specified, descriptive statistics for continuous data will include the number of subjects with data to be summarized (n), mean, standard deviation, median, and minimum and maximum. The same number of decimal places as in the raw data will be presented when reporting the minimum and maximum, one more decimal place than the raw data will be presented when reporting mean and median, and 2 more decimal places than the raw data will be presented when reporting standard deviation.

Descriptive statistics for categorical/qualitative data will include frequency counts and percentages. The total number of subjects in the treatment group will be used as the denominator for percent calculations, unless stated otherwise. All percentages will be presented with one decimal, unless otherwise specified. Percentages equal to 100 will be presented as 100, and percentages will not be presented for zero frequencies.

Descriptive statistics associated with time-to-event analyses will include the number of events, the number of subjects censored, 25% quartile, median, 75% quartile, and corresponding 95% confidence interval for median. These statistics will be presented for all time-to-event analyses, unless stated otherwise.

For listings broken down by center and treatment arm, site (center) number will be ordered by country.

6.2 Handling of Missing Data

In general, values for missing data will not be imputed unless methods for handling missing data are specified.

6.2.1 Handling of Missing Dates/Months/Years for Adverse Events

Adverse events (AEs) with incomplete onset dates will be handled as follows for the sole purpose of determining treatment emergence (TEAE is defined in Section 7.3.1):

- If the start/end date of an AE is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The AE will be assumed to be treatment emergent if it cannot be definitively shown that the AE did not occur or worsen during the treatment-emergent period (worst case approach). The detailed algorithm will be specified in ADaM specifications.
- If the start date is completely missing, an AE will be considered treatment-emergent unless the stop date is before study drug administration.
- If the dose start date is missing for a subject at a data-cut, all AEs of the subject will be considered treatment-emergent.

The original partial or missing date will be shown in all listings of AEs.

6.2.2 Handling of Missing Dates/Months/Years for Concomitant Medications

Prior or concomitant medications with incomplete start dates will be handled as follows for the sole purpose of determining whether a non-study medication is a concomitant medication:

- If the start/stop date of a medication is partially missing, the date will be compared as far as possible with the date of the start of administration of study drug and the date of last dose of study drug plus 30 days. The medication will be assumed to be concomitant if it cannot be definitively shown that the stop date is before the start of administration of study drug, or the start date is more than 30 days after the last date of administration of study drug. The detailed algorithm will be specified in ADaM specifications.
- If the start/stop dates are completely missing, a medication will be considered concomitant.
- If the dose start date is missing for a subject at a data-cut, all non-study medications of the subject will be considered concomitant.

The original partial or missing date will be shown in listings of all non-study medications.

6.2.3 Handling of Missing Dates/Months/Years for Disease History and Prior Therapies

For the purpose of calculating the duration from initial diagnosis, most recent relapse/refractory diagnosis or most recent prior therapy to randomization, partial/missing dates for diagnosis and last prior therapy completion will be imputed as follows:

- If both date and month are missing and the year is prior to the year of screening, the imputed date and month will be 01 July.
- If both date and month are missing and the year is the same as the year of screening, the imputed date will be the middle point between 01 Jan of the year and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- If date is missing and the month and year are prior to the month and year of screening, the imputed date will be 15th day of the month.
- If date is missing and the month and year are the same as the month and year of screening, the imputed date will be the middle point between the first date of the month and the screening date. If the middle point falls between two dates, the first of the two dates will be used.
- No imputation will be performed if the year is missing.

6.3 Multiple Comparisons/Multiplicity Adjustment

Of the secondary endpoints, ORR, Lymph Node Response Rate, OS, and Hematologic Improvement Rate are designated as key secondary efficacy endpoints. The primary endpoint and key secondary endpoints will be tested at an overall 1-sided alpha level of 0.025 based on a gatekeeping approach.

If the primary endpoint is significant, the 4 secondary endpoints will be sequentially tested at the 1-sided 0.025 significance level in the order listed above. If a null hypothesis is not rejected, formal sequential testing will be stopped.

6.4 Adjustments for Covariates

Adjustments for covariates will be considered for analysis of primary and key secondary endpoints, with details provided in Sections 7.2.1 and 7.2.2.

6.5 Subgroups

PFS, ORR and OS will be examined in the following subgroups:

- Stratification factors (captured on the eCRF):
    1 High-risk cytogenetics (presence vs absence of del[17p])
    2 Refractory/early relapse to purine analog-based therapy (defined as progression <12 months after fludarabine/pentostatin: yes vs no)
    3 Grade 4 cytopenia(s) (presence vs absence of neutropenia or thrombocytopenia at baseline)
- Diagnosis (CLL or SLL)
- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Previously treated with ofatumumab (yes vs no, if sample size allows the analysis)
- Most recent prior anti-cancer therapy to randomization <12 months: yes vs no

TEAEs (All Causalities) and TEAEs (Treatment-Related) will be examined in the following subgroups:

- Gender (Male or Female)
- Age group (<65 or ≥65)
- Race (White or Non-White)
- Baseline ECOG performance status (0 or 1 versus 2)

More details will be specified in Sections 7.2.1, 7.2.2 and 7.3.1.

6.6 Visit Windows

All data will be categorized based on the scheduled visit at which it was collected. These visit designators are predefined values that appear as part of the visit tab in the eCRF. There will be no additional analysis windowing done based on the assessment date.

6.7 Unscheduled Visits

Unscheduled visits will not be included in by-visit summary tables, unless otherwise specified. For laboratory tests, data from unscheduled visits will be included in listings and summaries of maximum changes from baseline, and the best or worst post-baseline values. For endpoints

based on disease status assessment, data from unscheduled assessments will be included in the derivation and analyses of the endpoints.

6.8 Baseline Values

Unless otherwise specified, the baseline value is defined as the value collected at the time closest to, and prior to, the start of study drug administration. Values collected at unscheduled visits prior to the start of the study drug administration will be included in the calculation of baseline values.

6.9 Computing and Coding Standards

Activities will be performed using the following tools:

Table, listing, and figure production	SAS Version 9.2 or higher
Coding	
Adverse Events	MedDRA Version 16.1 or higher
Medical Histories	MedDRA Version 16.1 or higher
Prior and Concomitant Medications	WHODrug Version September 2013
Grading	
AEs	CTCAE Version 4.03
Labs	CTCAE Version 4.03

7 STATISTICAL ANALYSES

7.1 Study Subjects

7.1.1 Disposition of Subjects

The disposition of subjects will include the number and percentage of subjects for the following categories: the number randomized, the number and percentage randomized but not dosed, the number and percentage dosed. These categories will be summarized for each treatment arm and for the two treatment arms combined (total). The percentages will be based on all randomized subjects (ITT analysis set).

An end-of-treatment disposition (still on treatment vs discontinued from treatment) will be provided for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set. The primary reason for treatment discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

An end-of-study disposition (still on study vs discontinued from study) will also be provided for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The primary reason for study discontinuation will be included in the table. A listing of these subjects will also be provided, broken down by center and treatment arm.

A summary of strata as captured in IRT will be presented by treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. A listing of the discrepancies in data for stratification factors between those captured in the IRT and those reported on the eCRF will be presented.

7.1.2 Protocol Deviations

Protocol deviations will be assessed during monitoring visits and by data review. At minimum, the following categories of study conduct or data will be assessed for potential protocol deviations: entry criteria, concomitant medications, dosing records, laboratory results, visit schedule and procedures. Protocol deviations will be categorized as major or minor prior to data release for the final analysis of the primary endpoint. A summary table of the protocol deviations will be provided by treatment arm and for two arms combined (total) for the ITT analysis set. A listing of all protocol deviations will be provided, broken down by center and treatment arm.

7.1.3 Demographic and Other Baseline Characteristics

Demographic variables will be summarized for each treatment arm based on all randomized subjects (ITT analysis set) and for the two treatment arms combined (total). The variables will include age, age group (<65 versus >=65), sex, race, ethnicity, height, and weight.

Demographic variables will also be summarized based on the PP analysis set.

7.1.4 Disease History

Disease history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include diagnosis (CLL or

SLL), risk factors used for randomization that are reported on eCRF (high risk cytogenetics [17p deletion: presence vs absence], refractory/early relapse to purine analog based treatment [yes vs no], grade 4 cytopenia(s) [presence vs absence]), years from initial diagnosis to randomization, months from most recent relapse/refractory diagnosis to randomization, stage at initial diagnosis, type of prior treatment, current stage, Binet/Rai stage, and baseline lymphocytes.

The durations to be summarized are defined as follows:

- Years from initial diagnosis to randomization will be calculated as (date of randomization – date of initial diagnosis + 1)/365.25.
- Months from most recent relapse/refractory diagnosis to randomization will be calculated as (date of randomization – date of most recent relapse/refractory diagnosis + 1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3.

7.1.5 Prior Therapies

Prior therapies will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT analysis set. The variables will include number of prior systemic therapies (summarized as a continuous variable and as a categorical variable), months from most recent prior therapy to randomization, number and percentage of subjects with prior radiotherapy, and number and percentage of subjects with prior surgery related to primary diagnosis.

The durations to be summarized are defined as follows.

- Months from most recent prior therapy to randomization will be calculated as (date of randomization – stop date of most recent systemic therapy, prior irradiation or prior surgery +1)/ (365.25/12).

The imputation of partial/missing dates is described in Section 6.2.3. A by-subject listing will be presented for prior systemic therapy.

7.1.6 Medical History

Medical history will be summarized for each treatment arm and for the two treatment arms combined (total) based on the AT analysis set.

Medical history will be summarized by system organ class (SOC) and preferred term (PT) using the number and percentage of subjects who had at least one occurrence of an SOC or PT. The summary will be sorted alphabetically in SOC and by decreasing frequency of PT in the IPI-145 arm within an SOC.

7.1.7 Prior and Concomitant Medications

Medications will be considered as prior if they stopped before the date of first dose of study drug.

Medications will be considered concomitant if they were taken at any time between the date of first dose of study drug and 30 days after the date of last dose of study drug, inclusive. If the start date or end date of a medication is completely or partially missing, refer to Section 6.2.2 for the algorithm to determine whether a medication is concomitant.

Prior medications and concomitant medications will be summarized separately. Both summaries will be based on the AT analysis set.

Medications will be summarized by ATC level 1, ATC level 2, and preferred drug name for each treatment arm. The summary will be sorted by decreasing frequency in ATC level 1, ATC level 2 and preferred drug name in the IPI-145 arm. A subject taking the same drug multiple times will only be counted once.

A listing will be provided for all non-study medications taken on the study. An identifier will be provided to show if a medication is prior or concomitant. Medications that started more than 30 days after the last dose of study drug will be identified as post-treatment.

7.1.8 Exposure to Study Drug

Extent of exposure will be summarized for each treatment arm based on the AT analysis set.

Extent of exposure will be summarized for the following variables:

- Duration (weeks): (date of last dose – date of first dose + 1) divided by 7
- Number of cycles started (continuous and categorical)
- Ofatumumab arm only: For each infusion (1-12), number and percent of subjects receiving the infusion
- Relative dose intensity, defined as 100% x (total dose received)/ (planned cumulative dose for the duration of treatment)
- Number and percentage of subjects with a dose reduction
- IPI-145 arm only: Number and percentage of subjects with a dose increase
- Number and percentage of subjects with a dose interruption
- Number and percentage of subjects with study drug discontinued

7.2 Efficacy Analyses

For PFS, ORR, lymph node response rate, DOR and other endpoints derived from progression and/or response status, the primary analyses will be based on the endpoints derived from independent central review (IRC). The endpoints derived from investigator assessment will be used in sensitivity analyses.

All efficacy analyses will be based on the ITT analysis set unless stated otherwise. If analyses are performed on more than one analysis set, the analyses on the ITT analysis set will be considered primary.

For stratified analysis of any efficacy endpoint, the following algorithm will be used to pool strata if there is insufficient information in any stratum (ie, there are <6 subjects, or there is no event for a time-to-event endpoint, or all subject have the same outcome for a binary endpoint in a stratum). (1) Strata will be ranked from the smallest to the largest based on the number of subjects, and if there is a tie, based on the number of events or responses. If there is still a tie, based on the reverse order of strata as determined by the stratification factors and levels (see Section 3.3). (2) The smallest stratum will be compared with the criteria of insufficient information. (3) If there is insufficient information in the smallest stratum, that stratum will be pooled with the smallest of the adjacent strata, which are defined as strata having 2 of the 3 stratification factors being at the same level as the smallest stratum. At the end of the three steps, the pooled stratum will replace the original two contributing strata. If there is still insufficient information in any stratum, the three steps will be repeated with the last pooled stratum assuming the stratum label of the larger of the two contributing strata for the purpose of additional pooling. The resulting pooled strata will be the strata used for the stratified analysis described in this section.

7.2.1 Analyses of Primary Endpoint

7.2.1.1 Primary analyses of PFS

Number of subjects with events, types of events (progression or death before progression), number of subjects censored, number of subjects for each reason of censoring, estimates and 95% confidence intervals for the 25th percentile, median, and 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points may also be presented.

A stratified log-rank test (1-sided) will be used to compare PFS of the IPI-145 arm against PFS of the ofatumumab arm at the interim and final analyses with the overall 1-sided significance level controlled at 0.025. The nominal p-value required for statistical significance at an analysis will be calculated based on the numbers of PFS events at all preceding and current analyses. The HR (IPI-145/ofatumumab) and the corresponding 2-sided 95% CI will be estimated using a stratified Cox proportional hazards model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

PFS will be plotted for each treatment group using the Kaplan-Meier method. The plot will include the number of subjects at risk for each treatment group over time.

7.2.1.2 Sensitivity analyses of PFS

The following sensitivity analyses will be performed:

- PFS based on investigator assessment: This endpoint will be analyzed using the same methods as the primary analyses described for PFS based on IRC assessment. In addition, the differences between investigator assessment and IRC assessment will be summarized.

- Worst-case sensitivity analysis: Subjects who are alive and have not had documented progression by data cutoff and who are "lost to follow-up" (missing at least one disease assessment right before data cut off) will be treated as censored at their last adequate disease assessment if they are on the control arm (ofatumumab) and treated as having a PFS event at the time of the next scheduled assessment following the last adequate disease assessment if they are on the experimental arm (IPI-145). PFS based on IRC assessment with the above worst-case censoring/event rule will be analyzed using the same methods as the primary analyses for PFS.
- Event-free survival (EFS): This is defined as time from randomization to the first documentation of PD as determined by IRC, start of new anticancer treatment or procedure, or death due to any cause. The event/censoring rule in Appendix A, which is used for the primary endpoint, will be used except that start of new anticancer treatment or procedure will be considered an event. EFS will be analyzed using the same methods as the primary analyses for PFS except that the summary of types of events will include a category of new anticancer treatment or procedure.
- Analysis using the PP analysis set.
- Analysis using the AT analysis set.
- Unstratified analyses: An unstratified log-rank test (1-sided) will be used to compare the two trreatment arms. An unstratified Cox regression (1-sided) will be used to estimate the hazard ratio (IPI-145/ofatumumab) with its 95% confidence interval. Other aspects of the analyses will be the same as the primary analysis of PFS.
- Cox regression with baseline covariates: A stratified Cox regression will be used to test treatment effect on PFS, adjusting for demographic and baseline characteristics. The strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. A stepwise variable selection will be performed to choose the variables in the Cox regression. Candidate variables are age, gender, race, disease diagnosis (CLL or SLL), years from initial diagnosis, months from most recent relapse/refractory diagnosis, stage at diagnosis, current stage, and number of prior systemic therapies. PFS based on IRC assessment will be used for this analysis.

7.2.1.3 Subgroup analyses of PFS

Subgroup analyses for PFS will be performed using the subgroups specified in Section 6.5. The HR (IPI-145/ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The hazard ratio and their 95% confidence intervals will be displayed for all subgroups graphically in a forest plot.

7.2.2 Analyses of Secondary Efficacy Endpoints

Of the secondary endpoints, ORR, Lymph Node Response Rate, OS, and Hematologic Improvement Rate are designated as key secondary efficacy endpoints. These endpoints will be tested for statistical significance under an overall 1-sided significance level of 0.025 if and only if the primary endpoint is significant. Details of multiplicity adjustment are provided in Section 6.3.

The analysis of another secondary efficacy endpoint, Duration of Response (DOR), will be descriptive only. No hypothesis testing will be performed.

7.2.2.1 Overall Response Rate

The primary analyses of ORR will use ORR based on IRC assessment in the ITT analysis set.

ORR will be derived from best overall response (BOR), which is defined as the best time point response that a subject achieves during the course of the study, with the response ranked according to the following order (from best to worst): CR>CRi>PR>PRwL>SD>PD (CRi applies to CLL only). In addition, the IRC may assign a timepoint response of unknown (UNK) due to missing, incomplete, or inadequate data, or a timepoint response of no evidence of disease (NED) if both radiological and clinical data indicate no disease involvement; for a post-baseline timepoint, the IRC may also assign a timepoint response of not evaluable (NE) if no target lesions were identified at baseline and the radiological and clinical data at the post-baseline timepoint do not support the disease response of PD or UNK; these categories will be classified as OTHER.

The estimated ORR (percent of subjects with a BOR of CR, CRi, PR or PRwL) and a 2-sided 95% CI will be provided for each treatment arm. Number and percent of subjects with BOR in each of the response categories (CR, CRi, PR, PRwL, SD, PD, OTHER) will also be presented by treatment arm. All subjects in the analysis set will be included in the denominator in the calculation of the percentage for each response category or ORR.

ORR will be analyzed using the Cochran-Mantel-Haenszel test (1-sided) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. The odds ratio and its 95% CI will be estimated.

The following sensitivity analyses will be performed:

- ORR based on investigator assessment
- Overall Confirmed Response Rate (OCRR), with overall confirmed response (based on independent review) defined as best confirmed response (≥8 weeks in duration) of CR, CRi, PR, or PRwL, according to the IWCLL or revised IWG Response Criteria, with modification for treatment-related lymphocytosis
- Overall Response Rate without PRwL: with overall response (based on independent review) defined as best response of complete response/remission (CR), CR with incomplete marrow recovery (CRi), partial response/remission (PR)
- Overall Confirmed Response Rate without PRwL
- Overall Response Rate in subset of subjects with baseline assessment other than UNK and NED. This analysis will be performed if there are at least 5% subjects with baseline assessment of UNK or NED.
- Analysis using the PP analysis set.
- Analysis using the AT analysis set.

Subgroup analyses will be performed for ORR using the subgroups specified in Section 6.5. ORR will be analyzed using the (unstratified) Chi-Square test to compare the two treatment groups in each subgroup. A forest plot of estimated odds ratios and their 95% confidence intervals will be presented. An estimated ORR and its 95% confidence interval in each subgroup will also be provided.

7.2.2.2 Lymph Node Response Rate

The primary analyses of lymph node response rate will be based on IRC assessment in the ITT analysis set. The number of subjects with lymph node response, estimated lymph node response rate, and a 2-sided 95% CI will be provided for each treatment arm.

Lymph node response rate will be analyzed using the Cochran-Mantel-Haenszel test (1-sided) to compare the two treatment groups. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

A sensitivity analysis will be performed using lymph node response rate based on investigator measurements.

7.2.2.3 Overall Survival

The primary analyses of OS will be based on the ITT analysis set. Subjects without documentation of death at the time of the data cutoff for analysis will be censored at the date the subject was last known to be alive, or the data cutoff date, whichever is earlier. A stratified 1-sided log-rank test will be used to compare OS between the 2 treatment groups. The HR along with the 95% CI will be estimated using a stratified Cox model. For both the stratified log-rank test and the stratified Cox model, the strata will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2. The Kaplan-Meier estimates of survival curves and 95% confidence intervals for the 25th percentile, median, and 75th percentile for OS will be presented by treatment group (if estimable). Probabilities of survival at selected time points may also be presented.

A majority of subjects can be expected to take subsequent anticancer therapy, especially after disease progression. To adjust for the effects of subsequent therapy, the following sensitivity analysis may be performed:

1. Primary analyses except additional censoring at the start day of subsequent therapy
2. Rank preserving structural failure time (RPSFT) analysis
3. Inverse probability of censoring weighted (IPCW) analysis

The following sensitivity analyses will be performed:

- Analysis using the PP analysis set.
- Analysis using the AT analysis set.

- Unstratified analyses: An unstratified log-rank test (1-sided) will be used to compare the two treatment arms. An unstratified Cox regression (1-sided) will be used to estimate the hazard ratio (IPI-145/ofatumumab) with its 95% confidence interval.

Subgroup analyses for OS will be performed using the subgroups specified in section 6.5. The HR (IPI-145/ofatumumab) with its 95% CI will be calculated based on an unstratified Cox regression model for each subgroup. The HRs and their 95% CIs will be displayed for all subgroups graphically in a forest plot.

7.2.2.4 Hematologic Improvement Rate

A subject with a hematologic improvement is one who consistently met the criteria of an improvement in neutrophil count, hemoglobin or platelet count for a period of at least 60 days during which the subject did not have a transfusion or exogenous cytokines. Missing data or missing scheduled assessments during a potential 60-day hematologic improvement episode will result in categorization of not having a hematologic improvement.

The number of subjects with hematologic improvement, estimated hematologic improvement rate, and a 2-sided 95% CI will be provided for each treatment arm.

Hematologic improvement rate will be compared between treatment groups using 1-sided Cochran–Mantel–Haenszel (CMH) test. The strata for the test will be those used for stratified randomization, with potential pooling of strata specified earlier in Section 7.2.

7.2.2.5 Duration of Response (DOR)

DOR will be presented using the Kaplan-Meier method for each treatment arm based on all randomized subjects with a documentation of response (i.e., CR, CRi, PR, or PRwL) as determined by IRC assessment. The censoring rule will be the same as that for the primary endpoint (see Appendix A). The Kaplan-Meier estimates of survival curves and 95% confidence intervals for the 25th percentile, median, and 75th percentile for DOR will be presented by treatment group (if estimable). No treatment comparison will be performed.

7.2.3 Analyses of Exploratory Efficacy Endpoints

7.2.3.1 Improvement in Disease-Related Symptoms

Disease-related symptoms of fever, weight loss, and drenching night sweats will be assessed as present or absent at each timepoint. Improvement in each of the three disease-related symptoms from baseline will be analyzed separately. The summary will display the proportion of subjects with an improvement while on treatment. Descriptive analysis of improvement in these disease-related symptoms will also be presented over the course of the study (eg, for each scheduled assessment).

The disease-related symptom of fatigue will be measured by the ECOG performance status. Summary statistics will be provided for the ECOG score and the change from baseline at each timepoint.

7.2.3.2 Minimal Residual Disease

MRD status for subjects with a response of CR or CRi will be determined using flow cytometry from a central lab. MRD will be classified as positive or negative. The estimated rates of subjects with negative MRD and their 95% confidence intervals in subjects with a response of CR or CRi will be provided for each treatment arm.

7.3 Safety Analyses

All safety analyses will be performed using the AT analysis set.

7.3.1 Adverse Events

Adverse events will be coded using MedDRA Version 16.1 or higher. The Grade of the AE will be assessed according to the NCI-CTCAE Version 4.03. If an AE is not included in the NCI-CTCAE Version 4.03, the Grade of the AE will be assessed according to the protocol, Section ***8.2.1.2***.

AEs will be summarized for each treatment arm by MedDRA system organ class (SOC) and preferred term (PT), or PT only. For summary tables by SOC and PT, SOC will be sorted alphabetically and PT will be sorted by decreasing frequency in the IPI-145 arm within each SOC. For summary tables by PT only, PT will be sorted by decreasing frequency in the IPI-145 arm.

If multiple AEs of the same PT occur within a subject, only the maximum grade observed for this PT will be used in summary of AEs by grade, the subject will be counted only once in the number of subjects for this PT and only once for the number of subjects for the SOC to which this PT belongs.

A treatment-emergent AE (TEAE) is defined as any AE that emerges or worsens in the period from the first dose of study treatment to 30 days after the last dose of study treatment. The onset date of an AE will be compared to the first dose date and the last dose date plus 30 days to determine whether the AE is treatment-emergent or not. If the onset date of an AE is completely or partially missing, refer to Section 6.2.1 for the algorithm to determine whether an AE is treatment emergent.

An overview TEAE summary table will be provided, which will include the number of AEs and the number of subjects with AEs in selected categories. In addition, TEAEs will be summarized for the following categories, and will be tabulated by SOC and PT, unless otherwise specified.

- Treatment-emergent AEs (All Causalities)
- Treatment-emergent AEs (Treatment-Related)
- Treatment-emergent AEs (All Causalities, by maximum grade)
- Treatment-emergent AEs (Treatment-Related, by maximum grade)
- Grade 3 or higher treatment-emergent AEs (All Causalities)
- Grade 3 or higher treatment-emergent AEs (Treatment-Related)

- Treatment-emergent SAE (All Causalities)
- Treatment-emergent SAE (Treatment-Related)
- Treatment-Emergent AEs Resulting in Discontinuation of Study Drug
- Treatment-Emergent AEs Resulting in dose hold or reduction (number of subjects with hold, reduction, hold or reduction will be displayed)
- Treatment-Emergent AEs Resulting in Death
- Treatment-Emergent AEs by PT (All Causalities)
- Grade 3 or higher treatment-emergent AEs by PT (All Causalities)

A by-subject listing of the following AE categories will be presented.

- All AEs (TEAEs will be flagged)
- All SAEs (TEAEs will be flagged)
- Treatment-Emergent AEs Resulting in Discontinuation of Study Drug
- Treatment-Emergent AEs Resulting in Death

Treatment-emergent AEs (All Causalities) and Treatment-emergent AEs (Treatment-Related) will also be tabulated in each subgroup specified in Section 6.5.

7.3.2 Laboratory Data

Laboratory tests will be reported separately for hematology/coagulation, blood chemistry, and urinalysis.

For the purposes of presentation in both tables and listings, the following laboratory test results will be converted to the International System of Units (SI) before presentation: sodium, potassium, chloride, bicarbonate (or CO_2), albumin, total protein, creatinine, uric acid, calcium, phosphorus, magnesium, glucose, total and direct bilirubin, and alkaline phosphatase, red blood cell (RBC) count, hemoglobin, hematocrit, platelets, white blood cell count with 5-part differential performed manually or by flow cytometry (for an absolute neutrophil count [ANC], lymphocyte count, neutrophils, monocytes, basophils, and eosinophils), etc.

If a laboratory test value is reported using a non-numeric qualifier (e.g., less than [<] a certain value, or greater than [>] a certain value), the given numeric value will be used in the summary statistics, ignoring the non-numeric qualifier.

For laboratory tests with NCI-CTCAE grades, a shift table from baseline grade to the maximum post-baseline grade will be provided. Laboratory tests with bi-directional grades (e.g., Hyperglycemia and Hypoglycemia) will be presented separately for each direction within the shift table.

Listings will be provided for all laboratory test results and for laboratory test results grade 3 and higher. A listing of subjects with ALT or AST >3xULN with simultaneous total bilirubin >2xULN will be presented, where ULN stands for upper limit of normal.

7.3.3 Vital Signs

The actual values of vital sign parameters, including temperature, heart rate, weight and systolic and diastolic blood pressure, will be presented in a by-subject listing.

7.3.4 Electrocardiogram (ECG)

The ECG categories will be summarized for each treatment arm for baseline and for each visit. The categories include normal, abnormal, NCS (abnormal but not clinically significant) and abnormal CS (abnormal and clinically significant).

7.3.5 Concomitant Medications and Procedures

Please refer to Section 7.1.6 for the definition and summary of concomitant medications.

Concomitant procedures will not be summarized. A by-subject listing will be presented.

7.4 Quality of Life Instruments

Analyses of the QoL instruments (EQ-5D and FACIT-F) will be performed on the ITT analysis set. Additional analyses may be considered as needed.

7.4.1 EQ-5D

The EQ-5D contains a descriptive system with one response for each of 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). The first response is coded as 1 (indicating no problems), the second response is coded as a 2 (indicating some problems), and the third response is coded as a 3 (indicating extreme problems). Ambiguous responses (eg, more than one response in a dimension) are treated as missing values. The EQ-5D also contains a Visual Analog Scale (VAS) for health state ranging from 0 to 100, where 0 represents worst imaginable health state and 100 represents the best imaginable health state.

For each dimension of the descriptive system, the number and percentage of subjects with no problems, some problems, and extreme problems will be reported at each visit.

Summary statistics for the VAS of health state and the change from baseline will be reported for each visit.

7.4.2 FACIT-F

The FACIT-F contains 5 subscales, Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, Functional Well-Being, and Fatigue (Additional Concerns on eCRF). Each subscale contains multiple items, and each item has a response of 0 to 4. Scores will be computed for each subscale, as well as for the FACIT-F Trial Outcome Index, FACT-G total score, and the FACIT-F total score.

The Physical Well-Being subscale contains 7 items. To compute the subscale score, compute the item score by subtracting each item response from 4 (item score = 4 – item response), sum the

item scores, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Physical Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Social/Family Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Social/Family Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Emotional Well-Being subscale contains 6 items. To compute the subscale score, compute the item score first. For items 1, 3, 4, 5, and 6, subtract each item response from 4 (item score = 4 – item response). For item 2, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 6 and divide by the number of items answered to compute the subscale score. At least 3 questions must be answered. If fewer than 3 questions are answered, then the subscale score will be missing. The Emotional Well-Being score ranges from 0 to 24 with higher scores representing better quality of life.

The Functional Well-Being score contains 7 questions. To compute the subscale score, sum the item responses, multiply the sum by 7 and divide by the number of items answered. At least 4 questions must be answered. If fewer than 4 questions are answered, then the subscale score will be missing. The Functional Well-Being score ranges from 0 to 28 with higher scores representing better quality of life.

The Fatigue subscale contains 13 items. To compute the subscale score, compute the item score first. For all items except An5 and An7, subtract each item response from 4 (item score = 4 – item response). For items An5 and An7, the item score is the same as the item response (no subtraction is needed). Sum the item scores, multiply the sum by 13 and divide by the number of items answered to compute the subscale score. At least 7 questions must be answered. If fewer than 7 questions are answered, then the subscale score will be missing. The Fatigue score ranges from 0 to 52 with higher scores representing better quality of life.

To derive the FACIT-F Trial Outcome Index, sum the subscale scores for Physical Well-Being, Functional Well-Being, and Fatigue. If any of these subscale scores is missing, the score for the FACIT-F Trial Outcome Index will be missing. The score ranges from 0 to 108.

To derive the FACT-G score, sum the subscale scores for Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, and Functional Well-Being. If any of these subscale scores is missing, the FACT-G score will be missing. The score ranges from 0 to 108.

To derive the FACIT-F total score, sum all 5 subscale scores. If any subscale score is missing, the FACIT-F total score will be missing. The score ranges from 0 to 160.

For each of the five subscale scores and for each of the FACIT-F Trial Outcome Index, the FACT-G score, and the FACIT-F total score, summary statistics for the actual values and changes from Baseline will be reported at each visit.

7.5 Pharmacokinetic Analyses

Plasma samples will be analyzed for IPI-145 and IPI-656 (metabolite) concentrations using a validated high performance liquid chromatography-tandem mass spectrometry (LC/MS/MS) method. The PK data collected will be analyzed by standard population PK methods, using appropriate software. Analysis of exposure-response relationships for efficacy and safety endpoints will be conducted. If there is only a limited amount of plasma concentration data from this study, the data may be pooled with the results of other studies to perform the population PK and exposure-response analyses. Further details on these analyses will be outlined in a separate analysis plan. Results of the population PK and exposure-response analysis will be summarized in a separate technical report.

7.6 Biomarker Analyses

The following exploratory analyses may or may not be presented formally. If presented, they will be limited to biomarkers for which data from a sufficient number of subjects are available.

7.6.1 Pharmacodynamic and Potentially Predictive Biomakers

The relationship between tumor genomics, gene expression, protein expression, and clinical endpoints will be explored as follows:

- Evaluation of serum biomarkers and blood immunophenotype and disease biomarkers for pharmacodynamic and predictive diagnostic relationships with IPI-145 clinical activity and safety
- Evaluation of tumor genomic features (eg, cytogenetics, FISH, DNA sequence variation, DNA copy number variation, and/or RNA expression) for pharmacodynamic and predictive diagnostic relationships with IPI-145 clinical activity
- Evaluation of the relationship between serum and/or tumor biomarkers and disease progression of IPI-145 and ofatumumab

7.6.2 Pharmacogenomics

DNA may be extracted from the pharmacogenomics sample in order to evaluate a subject's germline DNA. This sample may be used to explore the following:

- Evaluate the relationship between germline DNA sequence variations and PK of IPI-145
- Evaluate germline DNA as a control to verify that tumor sequence variations are somatic mutations

8 CHANGES IN PLANNED ANALYSES FROM PROTOCOL

None.

9 REFERENCES

1 O'Brien PC, Fleming TR. *A multiple testing procedure for clinical trials.* Biometrics 1979; 35:549-556.

2 Robins JM. *Information recovery and bias adjustment in proportional hazard regression analysis of randomized trials using surrogate markers*. Proceedings of the Biopharmaceutical Section, American Statistical Association 1993; pp. 24-33.

3 Robins JM, Finkelstein D. *Correcting for Non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests*. Biometrics, 2000; 56(3):779-788.

4 Robins JM, Tsiatis, AA. *Correcting for noncompliance in randomized trials using rank preserving structural failure time models.* Communications in Statistics, 1991; 20, 2609-2631.

10 APPENDICES

A. PFS censoring rule

Censoring of PFS will be performed as detailed in the table below.

Primary PFS Censoring / Event Methodology

Situation	Date of Event or Censoring	Outcome
No adequate baseline disease status assessment	Date of randomization	Censored
No adequate post-baseline disease status assessment unless death occurs prior to first post-baseline assessment	Date of randomization + 1	Censored
No documented progression or death before data cutoff	Date of last adequate disease status assessment	Censored
Documented progression with ≤1 missing scheduled disease status assessment before progression	Date of the earliest assessment that results in a finding of unequivocal progression	Event
Death before progression being documented with ≤1 missing scheduled disease status assessment before death	Date of death	Event
Documented progression or death following a long gap between adequate disease status assessments (eg, 2 or more consecutive missed scheduled disease status assessments)	Date of last adequate disease status assessment before the gap	Censored
New anticancer treatment or procedure started before documented progression	Date of last adequate disease status assessment	Censored

Note: Disease status assessment includes CT scans (chest, abdomen and pelvis), bone marrow aspirate and/or biopsy (may not be required of all subjects at all scheduled disease status assessments), CBC and differential count, focused physical examination, disease related constitutional symptoms for disease assessment, and ECOG performance status. An adequate disease status assessment is any disease status assessment for which the blinded independent central review is able to arrive at a disease status (eg, CR, CRi, PR, PRwL, SD, and PD) per protocol-defined criteria.

B. Cochran–Mantel–Haenszel (CMH) test

The Cochran–Mantel–Haenszel (CMH) test compares binary responses of two treatment groups, adjusting for stratification factors. In the CMH test, the data are arranged in a series of associated 2 × 2 contingency tables, the null hypothesis is that the observed response is independent of the treatment used in any 2 × 2 contingency table.

Let O_{hij} be the observed frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h and E_{hij} be the expected frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h, where

$$E_{hij} = \frac{(O_{hi1} + O_{hi2}) \cdot (O_{h1j} + O_{h2j})}{O_{h11} + O_{h12} + O_{h21} + O_{h22}}$$

Also, let P_{hij} be the probability of outcome being j given treatment being i (i=1 or 2, j= 1 or 2) in stratum h. To test the following hypothesis,

$$H_0 : P_{h11} = P_{h21} \text{ for all } h \in \{1, \ldots, H\}$$
versus
$$H_1 : P_{h11} > P_{h21} \text{ for at least one } h \in \{1, \ldots, H\}$$

The 1-sided Cochran-Mantel-Haenszel Test Statistics is constructed as:

$$Z_{CMH} = \frac{\sum_{h=1}^{H} (O_{h11} - E_{h11})}{\sqrt{V_{11}}} \sim Z$$

Where

$$V_{11} = Var(\sum_{h=1}^{H} (O_{h11} - E_{h11})) = \sum_{h=1}^{H} \frac{E_{h11} \bullet E_{h22}}{O_{h11} + O_{h12} + O_{h21} + O_{h22} - 1}$$

The test statistic Z_{CMH} will be compared with standard normal distribution to obtain p-value of the CMH test.