A STUDY OF NERATINIB PLUS CAPECITABINE VERSUS LAPATINIB PLUS CAPECITABINE IN PATIENTS WITH HER2+ METASTATIC BREAST CANCER WHO HAVE RECEIVED TWO OR MORE PRIOR HER2-DIRECTED REGIMENS IN THE METASTATIC SETTING (NALA)

NCT01808573

Statistical Analysis Plan v2.1 17 Jan 2018

Statistical Analysis Plan – Supplemental 14 Nov 2018

Statistical Analysis Plan

Protocol Title:	A Study of Neratinib Plus Capecitabine Versus Lapatinib Plus Capecitabine in Patients with HER2+ Metastatic Breast Cancer Who Have Received Two or More Prior HER2-Directed Regimens in the Metastatic Setting (NALA)
Study Protocol Number:	PUMA-NER-1301
Disease Condition:	HER2+ Metastatic Breast Cancer
Sponsor:	Puma Biotechnology, Inc. 10880 Wilshire Blvd, Suite 2150 Los Angeles, CA 90024 USA Phone: +1 424.248.6500 Fax: +1 424.248.6501
Prepared by:	Puma Biotechnology, Inc. 10880 Wilshire Blvd, Suite 2150 Los Angeles, CA 90024 USA Phone: +1 424.248.6500 Fax: +1 424.248.6501

Date of Plan:	Original v1.1	29 Dec 2012
	Version 2.0	13 Nov 2017
	Version 2.1	17 Jan 2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Term
AE	adverse event
BID	twice daily
BMI	body mass index
CB	clinical benefit
cm	centimeters
CMH	Cochran-Mantel-Haenzel
CNS	Central Nervous System
CR	complete response
CRF	case report form
CSR	Clinical Study Report
CTCAE	Common Terminology Criteria for Adverse Events
DOR	duration of response
ECHO	echocardiogram
ECOG	Eastern Cooperative Oncology Group
EKG	electrocardiogram
EORTC QLQ-C30	European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire
EORTC QLQ-BR23	European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire Breast Cancer Module
EOS	end of study
EQ-5D-5L	EuroQol questionnaire
EQ-VAS	Visual Analog Scale
HER	human epidermal growth factor receptor
HER2+	human epidermal growth factor receptor 2 positive
IDMC	Independent Data Monitoring Committee
ITT	intent to treat
kg	kilogram
LVEF	left ventricular ejection fraction
m	meter
MBC	metastatic breast cancer

Abbreviation	Term
MedDRA	Medical Dictionary for Regulatory Activities
mg	milligram
mmHg	millimeters of mercury
MUGA	multiple-gated acquisition scan
ORR	objective response rate
OS	overall survival
PD	progressive disease
PFS	progression-free survival
PK	pharmacokinetics
PR	partial response
PT	preferred term
QoL	Quality of Life
RECIST	Response Evaluation Criteria in Solid Tumors
SAE	serious adverse event
SAP	Statistical Analysis Plan
SD	stable disease
SOC	System Organ Class
std	standard deviation
TEAE	treatment-emergent adverse event
WHODrug	World Health Organization Drug Reference List

1. PURPOSE OF THE ANALYSES

This statistical analysis plan (SAP) provides a comprehensive and detailed description of strategy and statistical methodology to be used for analysis of data for the PUMA-NER-1301 study. This study is sponsored by Puma Biotechnology, Inc.

The purpose of the SAP is to ensure the credibility of the study findings by pre-specifying the statistical approaches to the analysis of study data prior to data base lock. This analysis plan is meant to supplement the study protocol. Any deviations from this plan will be described in the Clinical Study Report (CSR).

2. PROTOCOL SUMMARY

2.1. Study Objectives

2.1.1. Primary Objective

The co-primary objectives of this study are

- to compare independently adjudicated progression free survival (PFS) following treatment with neratinib plus capecitabine versus lapatinib plus capecitabine in patients with HER2-positive (HER2+) metastatic breast cancer (MBC) who have received two or more prior HER2 directed regimens in the metastatic setting.
- to compare overall survival (OS) following treatment with neratinib plus capecitabine versus lapatinib plus capecitabine in this population.

2.1.2. Secondary Objectives

The secondary objectives of this study are:

- Investigator assessed PFS.
- Objective response rate (ORR), duration of response (DOR) and clinical benefit rate (CBR) (complete response [CR] or partial response [PR] or stable disease [SD] ≥24 weeks).
- Time to intervention for symptomatic metastatic central nervous system (CNS) disease.
- Safety (adverse events [AEs] and serious adverse events [SAEs]).
- Health outcomes assessment.

2.1.3. Exploratory Objective

The exploratory objective of this study is to assess the population pharmacokinetics (PK) of neratinib when administered in combination with capecitabine.

2.2. Overall Study Design and Plan

This is a randomized, multi-center, multinational, open-label active-controlled, parallel design study of the combination of neratinib plus capecitabine versus the combination of lapatinib plus capecitabine in HER2+MBC patients, who have received two or more prior HER2-directed regimens in the metastatic setting. Patients will be randomized in a 1:1 ratio to either the neratinib plus capecitabine or lapatinib plus capecitabine arms.

Following a 21-day screening period, eligible patients will be randomized to Arm A or Arm B. Baseline assessments will be performed prior to randomization and Cycle 1/Day 1 dosing. Patients will then participate in the active treatment phase, consisting of 21-day treatment cycles. Patients are anticipated to participate in the study for an average of 28 months. This includes approximately 0.5 months for screening, an estimated average of 9.5 months for treatment, and

an estimated average of 18 months for the long-term follow-up phase. Patients who permanently discontinue treatment will enter the long-term follow-up phase until death or withdrawal of consent.

2.3. Study Population

The study population will include all patients that sign and date an approved informed consent form and are eligible per the criteria defined in the protocol. A full list of the inclusion and exclusion criteria can be found in the PUMA-NER-1301 protocol.

2.4. Treatment Regimens

Patients in the Neratinib + Capecitabine Combination arm will receive medication as described below:

- Neratinib: six 40 mg tablets (total daily dose 240 mg) orally, once daily with food, preferably in the morning, continuously.
- Capecitabine: 150 mg or 500 mg tablets (total dose of 1500 mg/m^2 daily in 2 evenly divided doses), orally with water within 30 minutes after a meal. Dose taken daily for Days 1 to 14 of a 21-day cycle.

Patients in the Lapatinib + Capecitabine Combination arm will receive medication as described below:

- Lapatinib: five 250 mg tablets (total dose 1250 mg) orally, once daily, 1 hour before or after breakfast, continuously.
- Capecitabine: 150 mg or 500 mg tablets (total dose of 2000 mg/m^2 daily, in 2 evenly divided doses), orally with water within 30 minutes after a meal. Dose taken daily for Days 1 to 14 of a 21-day cycle.

2.5. Treatment Group Assignments or Randomization

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

- Arm A: neratinib (240 mg) + capecitabine (1,500 mg/m^2 daily, 750 mg/m^2 BID)
- Arm B: lapatinib (1250 mg) + capecitabine (2,000 mg/m^2 daily, 1000 mg/m^2 BID)

Patient randomization will be stratified according to:

- Number of previous HER2 directed regimens in the metastatic setting (N= 2 or N $\geq$3).
- Geographic region (North America vs. Europe including Russia and Israel vs. Rest of the World)
- Visceral vs. Non-Visceral only disease
- Estrogen receptor+ (ER+) and/or progesterone receptor (PR+) (i.e., hormone receptor +) versus ER- and PR- (i.e., hormone receptor -).

2.6. Sample Size Determination

The co-primary endpoints will be analyzed using an overall Type I error rate of 0.01 for PFS and 0.04 for OS. To detect a hazard ratio (experimental vs. control) of 0.70 with approximately 85% power, 419 PFS events (PD/Deaths) are required to be observed; this assumes a median (ITT) PFS of 8.0 months (34.7 weeks) for the experimental arm (neratinib plus capecitabine) and 5.6 months (24.3 weeks) for the control arm (lapatinib plus capecitabine).

It is expected that the 419th event will be observed at approximately 26 months after the first subject is enrolled, which would correspond to approximately half of the death events needed to be observed. At this time, an interim analysis on OS will also be performed. For the OS endpoint, a group sequential method will be employed to maintain the overall Type I error rate at 0.04. Specifically, Lan DeMets alpha spending function approximating the O'Brian-Fleming boundaries will be calculated to create appropriate critical values to which the log rank statistic will be compared (De Mets and Gordon Lan, 1994). If 50% of the deaths are observed at the interim OS analysis, the boundary using the Lan-DeMets alpha spending function is shown in Table 1 for OS interim analysis.

Table 1: Spending Function

Information Fraction	Expected Death Events	Study Month	Boundary (Reject H_o)	2-Sided Significance Level*
50%	185	26	3.066	0.002
100%	378	50	2.058	0.039

* This significance level relates to the 0.04 alpha allocation to OS but will be updated if the entire 0.05 can be allocated to OS.

These boundaries are designed to detect a hazard ratio (experimental vs. control) of 0.725 with approximately 85% power. Overall, 378 OS events (deaths) are required to be observed; this assumes a median (intent to treat [ITT]) OS of 30.3 months for the experimental arm (neratinib plus capecitabine) and 22.0 months for the control arm (lapatinib plus capecitabine).

Approximately 600 subjects will be enrolled and randomized equally between the two treatment arms. This sample size assumes a 1% drop out rate per 12 months (52 weeks) equally distributed between the treatment arms for the PFS endpoint.

3. GENERAL ANALYSIS AND REPORTING CONVENTIONS

All study variables will be listed. Selected variables will be summarized and, when appropriate, analyzed. Data summaries will be presented by treatment group and, when appropriate, visit.

For all safety and efficacy variables, the number of patients will be provided overall and by visit, with the exception of adverse events and concomitant medications, which will not be summarized by visit.

Categorical variables will be summarized using counts and percents. Percents will be displayed to 1 place after the decimal point (XX.X), with the exception of 100%, which will be displayed without additional decimal places. Continuous variables will be summarized using number of patients, mean, median, standard deviation, minimum, and maximum.

Mean and median will be reported at 1 more significant digit than the precision of the data; standard deviation will be reported at 2 more significant digits than the precision of the data. Minimum and maximum will be reported to the same level of precision as the original observations. In general, any calculated values, such as those due to unit conversion, will be rounded to the same number of decimal places as the original data. P-values will be reported to 3 decimal places, with values less than 0.001 displayed as '<0.001.'

In general, the baseline value will be considered the last measurement observed prior to taking the first dose of study treatment.

Data will be listed by treatment and patient. In general, listings will be sorted in the order that columns are displayed, starting with the first column on the left (treatment).

SAS statistical software, version 9.1 or later, will be used for all analyses.

If departures from these general conventions are present in the specific evaluations section of this SAP, then those conventions will take precedence over these general conventions.

4. ANALYSIS POPULATIONS

Four analysis populations will be used for this study: the Intent to treat (ITT) population, the Safety Population, the Pharmacokinetics (PK) population and the Quality of Life population (QoL).

The analysis of the primary and secondary efficacy endpoints will be performed using the ITT population. All safety analyses will be performed using the Safety Population. Exploratory Pharmacokinetics analyses will be performed using the PK population. Quality of life analyses will be performed using the QoL population.

4.1. Intent to treat (ITT) Population

The intent to treat population is defined as all patients who are randomized into the study. Patients will be analyzed in the treatment arm to which they were randomly assigned regardless of which treatment they received.

4.2. Safety Population

The safety population is defined as all patients who received at least one dose of investigational product. Patients will be analyzed based on the treatment they had received regardless of the treatment to which they were randomized.

4.3. Pharmacokinetics (PK) Population

The pharmacokinetics population consists of all patients for whom data on concentrations of study treatment is available.

4.4. Quality of Life (QoL) Population

The QoL population consists of all patients in the safety population for whom a baseline and at least one post baseline Health Outcomes Assessment (HOA) are available.

5. STUDY PATIENTS

5.1. Disposition of Patients

The number and percentage of patients entering and completing each phase of the study will be presented, stratified by treatment. Reasons for withdrawal pre- and post-randomization will also be summarized.

5.2. Protocol Deviations

Protocol deviations will be classified and monitored regularly during the duration of the study. Among other reasons, failure to meet any of the protocol inclusion or exclusion criteria will be considered a protocol deviation. All protocol deviations will be listed and summarized by type and treatment group.

6. DEMOGRAPHIC AND OTHER BASELINE CHARACTERISTICS

Descriptive statistics will be provided for all demographic and baseline characteristics based on the ITT population. Demographic data, medical history, concomitant disease, and concomitant medication will be summarized by means of descriptive statistics or frequency tables, overall and stratified by treatment. For categorical variables, the number and percentages of patients in each category will be presented. For continuous variables, summaries will include the number of patients with data, mean, median, standard deviation, minimum, and maximum.

6.1. Demographic and baseline characteristics

The following demographic and baseline variables will be summarized by treatment group:

- age (years)
- age group (<65, ≥ 65 years)
- sex
- race & ethnicity (per CRF)
- region
- height (cm)
- weight (kg)
- body mass index (BMI)

  Body mass index (BMI) will be calculated as:

  BMI (kg/m^2) = Weight (kg)/ (Height(cm) * 0.01)^2

6.2. Medical History

Medical history data including: chronic conditions, relevant surgical procedures, symptoms experienced during the previous 30 days, symptoms ongoing at the time of Screening, any

medical conditions that require medication and cancer history will be collected at screening, within 21 days before Cycle 1/ Day1 in accordance with the Schedule of Procedures included in the protocol.

Cancer history variables include date of first diagnosis, nodal status, histology, tumor stage at diagnosis, previous chemotherapy/biotherapy/immunotherapy, previous adjuvant therapy, previous radiation, and prior cancer related surgical therapies.

Medical history and cancer history data will be summarized and listed.

6.3. Prior and Concomitant Medications

Concomitant medications will be defined as medications documented on the Concomitant Medications CRF. Concomitant medications will be coded using the World Health Organization (WHODrug) dictionary and summarized in a table and a data listing.

7. MEASUREMENTS OF TREATMENT COMPLIANCE

Duration of treatment will be summarized separately for each study compound by treatment group. In addition, the cumulative quantity, dose intensity (quantity per time unit), and the relative dose intensity (dose intensity/scheduled dose per time unit) will be summarized.

The number of patients with dose holds or dose reductions will be tabulated for each test article by treatment group.

8. EFFICACY EVALUATION

8.1. Overview of Efficacy Analysis Issues

8.1.1. Handling of Dropouts or Missing Data

Missing assessments for PFS, the number of subjects who drop out prior to PFS or OS, and the length of follow up for dropouts will be summarized by treatment group. Sensitivity analyses will be performed to assess the impact of subject dropout or missing assessments on the primary efficacy analyses.

Other endpoints that could be affected by missing values (e.g. health outcome assessments) will be analyzed using a sensitivity analysis to quantify the impact of missing values in the results.

8.1.2. Multicenter Studies

Data from all sites will be pooled for the purpose of analyses. Exploratory analyses by site are not an objective of this study and are not currently planned.

8.2. Efficacy Variables

Table 2 gives an overview of the efficacy variables and the analysis methods that will be used for analysis.

Table 2: Efficacy Variables and Analysis Methods

Endpoint	Method
Primary	
Progression Free Survival (independently adjudicated)	• Kaplan-Meier estimates • Stratified Log-rank test • Stratified Cox Proportional Hazards model
Overall Survival	• Kaplan-Meier estimates • Stratified Log-rank test • Stratified Cox Proportional Hazards model
Secondary	
Progression Free Survival (investigator assessed)	• Kaplan-Meier estimates • Stratified Log rank test • Stratified Cox Proportional Hazards model
Objective Response Rate (independently adjudicated and investigator assessed)	• Cochran-Mantel-Haenszel test (CMH)

Table 2: Efficacy Variables and Analysis Methods (Continued)

Endpoint	Method
Duration of Response (independently adjudicated and investigator assessed)	• Kaplan-Meier estimates
Clinical Benefit Rate (independently adjudicated and investigator assessed)	• Cochran-Mantel-Haenszel test (CMH)
Time to intervention for symptomatic metastatic central nervous system disease (CNS) (investigator assessed)	• Cumulative incidence function for time to intervention for symptomatic CNS
Health Outcome Assessments	• Random Effects Mixed Model

These are definitions of the criteria used to determine objective tumor response for target lesions. These components make up the endpoints which are defined below, see the protocol for more detail.

Complete Response (CR): Disappearance of all target and non-target lesions. Any pathological lymph nodes (whether target or non-target) must have reduction in short axis to <10 mm. All lymph nodes must be non-pathological in size (<10 mm short axis).

Partial Response (PR): At least a 30% decrease in the sum of diameters of target lesions, taking as reference the Baseline sum diameters.

Progressive Disease (PD): At least a 20% increase in the sum of diameters of target lesions, taking as reference the smallest sum on study (this includes the Baseline sum if that is the smallest on study). In addition to the relative increase of 20%, the sum must also demonstrate an absolute increase of at least 5 mm. PD can also be determined if there is unequivocal progression of existing non-target lesions. (Note: the appearance of one or more new lesions is also considered progression).

Stable Disease (SD): Neither, sufficient shrinkage to qualify for PR, nor sufficient increase to qualify for PD, taking as reference the smallest sum diameters while on study.

8.2.1. Progression Free Survival (PFS)

Progression free survival (PFS) is defined as the time interval from the date of randomization until the first date on which recurrence, progression (per RECIST v1.1), or death due to any cause, is documented. It is not necessary to confirm disease progression.

All tumor assessments and deaths collected will be used except in the following 2 cases:

1. If a subject misses 2 or more consecutive assessments immediately prior to the documentation of PD or death the subject will be censored at the last assessment prior to the missed visits (following the FDA guidance on clinical trial endpoints for the approval of cancer drugs and biologics).
2. For subjects that started a new anti-cancer therapy, all tumor assessments and deaths after the start date of new anti-cancer therapy will be excluded from the analysis.

After taking into account the two conditions listed above, subjects without documented PD or death, including those who dropped out, will be censored at their last tumor assessment. Patients who did not have any tumor assessments after the screening visit will be censored on the randomization date.

Assessment of PFS will be performed by a blinded, independent, central review of tumor assessments for all patients at Screening, and then after every 6 weeks from first dose of investigational product, regardless of treatment schedule modification (e.g., dose delay), until documented disease progression or death. PFS will be assessed using the Response Evaluation Criteria in Solid Tumors (RECIST v1.1).

8.2.2. Overall Survival (OS)

Overall survival is defined as the time from randomization to death due to any cause, censored at the last date known alive on or prior to the data cutoff employed for the analysis. Patients who switch anticancer therapy will continue to be followed for the OS assessment.

Survival data will be collected throughout the study treatment phase and during post-progression follow-up. Survival follow-up after patient discontinuation of investigational product will be conducted approximately 4 weeks after treatment discontinuation and then every 12 weeks to assess for survival until subject's loss to follow-up or death.

8.2.3. Central Nervous System Disease (CNS)

Time to intervention for symptomatic metastatic central nervous system disease is defined as the time from randomization to the first start date of an intervention for symptomatic metastatic CNS disease. Subjects that do not have an intervention for symptomatic metastatic CNS and do not die will be censored at the last date known alive on or prior to the data cutoff. Deaths are treated as competing events. Interventions for symptomatic CNS disease will include anti-cancer medication, cancer-related radiation therapy, cancer-related surgery/procedure, or concomitant medication/therapy as collected on the CNS CRF.

8.2.4. Objective Response Rate (ORR)

Objective response rate is defined as the proportion of patients demonstrating an objective response during the study. Objective response includes confirmed complete responses (CR) and partial responses (PR) as defined in the RECIST criteria included in the study protocol. Tumor assessments after the initiation of new anticancer therapy will not be included when determining ORR.

8.2.5. Duration of Response (DOR)

Duration of response is measured from the time at which measurement criteria are first met for CR or PR (whichever status is recorded first) until the first date of recurrence or progressive disease (PD) or death is objectively documented, taking as a reference for PD the smallest measurements recorded since enrollment, per RECIST v1.1. This value is censored at the last valid tumor assessment if PD or death has not been documented. Tumor assessments after the initiation of new anticancer therapy will not be included when determining DOR.

8.2.6. Clinical Benefit Rate (CBR)

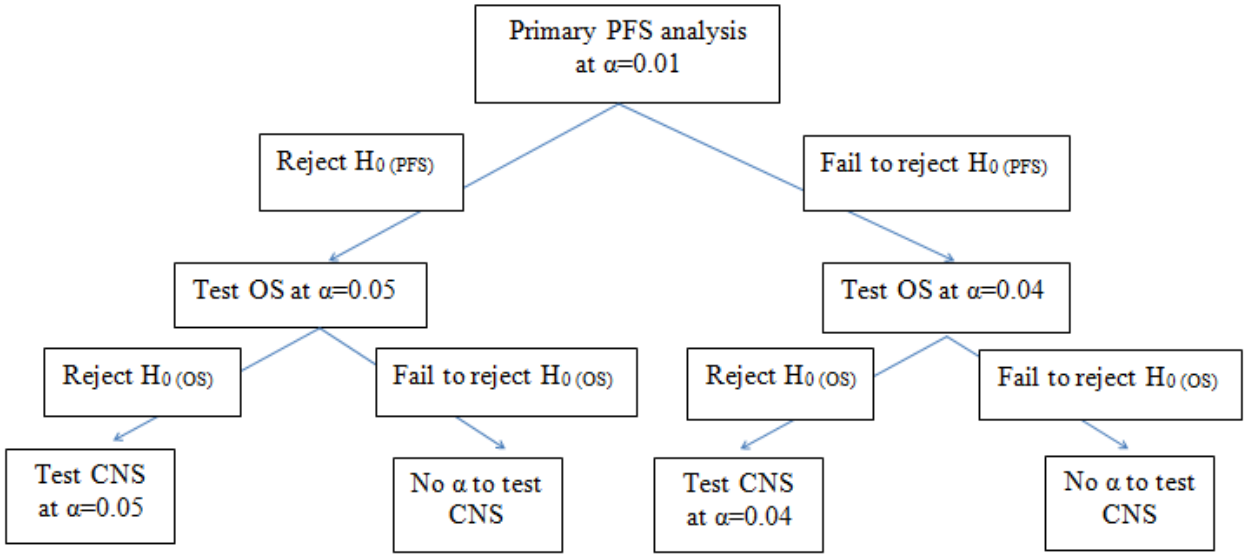
Clinical benefit rate is the proportion of patients who achieve overall tumor response (confirmed CR or PR) or stable disease (SD) lasting for at least 24 weeks from randomization. Tumor assessments after the initiation of new anticancer therapy will not be included when determining CBR.

8.3. Analysis Methods

The significance levels allocated to the co-primary efficacy endpoints are 0.01 for PFS and 0.04 for OS. However, once the null hypothesis for PFS is rejected at alpha = 0.01 at the time of the primary analysis then OS can be tested at the overall significance level of 0.05. The additional alpha propagated to the OS endpoint will not be spent at the interim OS analysis (Ye, 2013). The interim boundary for OS will remain the same as the boundary using the Lan-DeMets alpha spending function with overall alpha=0.04 based on the actual information fraction (number of deaths as a proportion of the 378 target number of deaths). The time to CNS intervention will be tested after PFS and OS. If PFS is not significant and OS is, then time to CNS will be tested at alpha =0.04; if both are significant, then time to CNS will be tested at alpha=0.05. Unless OS is significant at the interim analysis, time to CNS intervention will not be tested at the primary PFS analysis (interim OS), and will only be tested after OS is significant at the final OS analysis (Figure 1).

Test of all other secondary endpoints are descriptive.

Figure 1: Testing Schema



8.3.1. Primary Efficacy Analyses

The co-primary endpoints are OS and independently-assessed PFS. PFS will be determined by adjudicated disease progression (disease progression demonstrated by radiographic imaging or other appropriate modality), or death due to any cause, described in Section 8.2.1. Subjects will be stratified at randomization by whether they had 2 or ≥3 prior HER2+ MBC regimens in the metastatic setting; whether they had visceral disease or non-visceral only; their tumor's hormone receptor status, and their geographic region.

When 419 centrally assessed PFS events have occurred the primary PFS analysis will be performed. The median time to PFS and corresponding two-sided 95% confidence intervals will be calculated using the product limit estimator and displayed using a Kaplan-Meier graph, by treatment group. A log-rank test controlling for hormone receptor status, number of prior HER2-directed regimens in the metastatic setting, visceral disease status, and geographic region will be used to test the null hypothesis of no difference in the time to PFS between the two treatment groups at the alpha level of 0.01. A stratified Cox Proportional hazard model will be used to estimate the hazard ratio. PFS events will continue to be captured after the primary PFS analysis and a secondary, descriptive, analysis similar to the primary will be performed at the time of the primary OS analysis. When the PFS analysis is performed, an interim analysis will be performed on the OS endpoint. When 378 deaths have occurred, the primary (final) OS analysis will be performed. The median time to OS and corresponding two-sided 95% confidence intervals will be calculated using the product limit estimator and displayed using a Kaplan-Meier graph, by treatment group. A log-rank test controlling for hormone receptor status, number of prior HER2-directed regimens in the metastatic setting, visceral disease status, and geographic region strata will be used to test the null hypothesis of no difference in the survival between the two treatment groups as described in the overview of Section 8.3.

The co-primary efficacy analyses will be performed on the intent to treat (ITT) population defined as all subjects randomized into the study. Sensitivity analyses will be performed to

investigate the impact of dropout between the two treatment groups on the treatment effect estimate.

8.3.1.1. Sensitivity Analyses

The following sensitivity analyses for PFS will be performed:

1. If there are more than 10% of patients that start a new anti-cancer therapy prior to the documentation of a PD, the initiation of the new therapy will be considered as an event and the start date for the new anti-cancer therapy will be considered the PFS event date.
2. Similar to the primary analysis, a sensitivity analysis will be done without stratifying by the randomization strata. Identical event and censoring rules will be used as described in the primary PFS analysis (Section 8.2.1).
3. A sensitivity analysis will be performed using all tumor assessments and deaths. This means that new anti-cancer therapy will not be used to exclude tumor assessments and deaths and tumor assessments will be used even if a subject misses 2 or more visits prior to progression being documented.
4. Similar to the primary analysis, a sensitivity analysis will be done using interval censoring estimation to address any potential imbalance in intermittent missing assessments.

For OS, an unadjusted analysis without stratifying by the randomization strata will be performed.

8.3.2. Secondary Efficacy Analyses

The secondary endpoints of ORR, CBR and duration of response will be assessed using the centrally reviewed tumor assessments as well as the investigator tumor assessments. The secondary endpoint of time to intervention for symptomatic CNS disease will only use the investigator-reported data.

The treatment difference in the secondary efficacy endpoint of investigator assessed PFS will be assessed with the product limit estimate of the median time to event. Differences between treatment groups will be examined using a log rank test statistic stratified by hormone receptor status, number of prior HER2-directed regimen in the metastatic setting, visceral disease status, and geographic region strata (similar to the analyses for the primary endpoints). All time-to-event endpoints including time to intervention for symptomatic CNS disease will be displayed using a Kaplan-Meier survival plot.

For the secondary endpoints of ORR and CBR the subjects included in the analyses are those that had measurable disease (per RECIST v1.1) at screening. Any tumor assessments after the start of new anti-cancer therapy will not be included in the analyses for these two endpoints. The treatment difference in the secondary efficacy endpoints of ORR and CBR will be analyzed using a Cochran-Mantel-Haenzel (CMH) Chi-square test, stratified on hormone receptor status, number of prior HER2-directed regimens in the metastatic setting, geographic region and disease category (visceral vs. non-visceral only).

Duration of response will be summarized by treatment group for the patients that responded. If there are patients that continue to respond, they will be censored at their last tumor assessment and a Kaplan-Meier analysis will be done to estimate the duration of response.

With the exception of the CNS endpoint, which will be tested only after the null hypothesis is rejected for OS; all other secondary endpoints are descriptive analyses and will be done at the time of the primary PFS analysis.

Cumulative incidence of time to intervention for symptomatic CNS is defined as time from randomization to the start of therapy for symptomatic CNS. Competing event are deaths prior to intervention for symptomatic CNS. Gray's test (Gray, 1988) stratified for factors used in the randomization will assess the equality of cumulative incidence functions between the treatment arms. Additionally, the Kaplan Meier method, stratified log-rank test, and Cox proportional hazard model will be performed without considering competing events.

As a sensitivity analysis to the secondary CNS endpoint, the time to progressive central nervous system (CNS) lesions will be analyzed using the centrally reviewed tumor assessments as well as the investigator tumor assessments. Time to progressive CNS lesions is defined as the time interval from the date of randomization until the first date on which a scheduled or unscheduled visit documents the appearance of newly diagnosed or progressive CNS lesions, per RECIST v1.1. If there is no evidence of CNS progression, the subject will be censored at the last assessable evaluation on study or prior to new anti-cancer therapy, if applicable. Competing events include progression at other sites and deaths prior to progressive CNS. Gray's test (Gray, 1988) stratified for factors used in the randomization will assess the equality of cumulative incidence functions between the treatment arms. Additionally, the Kaplan Meier method, stratified log-rank test, and Cox proportional hazard model will be performed without considering competing events.

8.3.3. Other Efficacy Analyses

8.3.3.1. Health Outcome Assessments

The following health outcomes endpoints will be assessed at Cycle 1 Day 1 of the study (at any time within 24 hours of randomization) and at the end of every 6 weeks (± 3 days) starting from Cycle 1 Day1 as indicated in the Schedule of Procedures included in the protocol.

- European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire C-30 (EORTC QLQ-C30), version 3
- EORTC QLQ Breast Cancer Module (BR23)
- The EuroQol (EQ-5D) multi-dimensional health status questionnaire

8.3.3.1.1. EORTC QLQ C30 and BR23

The EORTC QLQ-C30 is a questionnaire developed to assess the quality of life of cancer patients which includes measures of physical functioning [PF], role functioning [RF], emotional functioning [EF], social functioning [SF], global QoL, and fatigue [FA]). It is supplemented by the breast cancer module (BR-23) which includes measures of body image, sexual functioning, systemic therapy side effects, breast symptoms, and arm symptoms. The complete questionnaire includes 53 items; 23 of them are specific of breast cancer. Patients will be asked to indicate the extent to which they have experienced symptoms or problems using a 4 point scale as follows: 1, is not at all; 2, a little; 3, quite a bit and 4, very much. Patients are also asked to rate their overall

health and quality of life during the past week using a 7 point scale where 1 is very poor and 7 is excellent.

Patient responses on the level of severity on the EORTC QLQ C30 and BR23 dimensions will be used to derive the corresponding EORTC QLQ C30 and BR23 scores, as per published methodology.

8.3.3.1.2. EQ-5D-5L

The EuroQol questionnaire (EQ-5D-5L) is a brief, self-administered generic health status instrument consisting of two parts. In the first part respondents are asked to describe their current health state on 5 dimensions (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression) with each dimension having 5 levels of function (1 = no problem, 2 = slight problem, 3 = moderate problem, 4 = severe problem, and 5= unable to function/extremely affected). The second part is a patient's self-rating of current health state on a Visual Analog Scale (EQ-VAS) with endpoints labeled 'best imaginable health state' and 'worst imaginable health state'. The EQ-5D-5L endpoints are the EQ-5D-5L Index which is derived by combining one level from each of the 5 dimensions and converting it to a single summary index or health utility value (Shaw et al., 2005) and the EQ-VAS score, which ranges from 0 for worst imaginable health state to 100 for best imaginable health state. Overall scores for the EQ-5D-5L index range from -0.594 to 1, with low scores representing a higher level of dysfunction.

Patient responses on the level of severity on the 5 EQ-5D-5L dimensions will be used to derive the corresponding EQ-5D-5L utility values, as per published methodology.

8.3.3.2. Analysis of Health Outcome Assessments

Patient-reported health-related quality of life measures (EORTC QLQ-C30 and EQ-5D) will be summarized by treatment group and visit. Differences between treatment groups will be described using an F test from a random effects mixed model with treatment, time, number of prior HER2-directed regimens in the metastatic setting (binary), geographic region, hormone receptor status, disease category, and the treatment by time interaction.

Endpoints that could be affected by missing values (e.g., health outcomes assessments) will be analyzed using a sensitivity analysis, including pattern mixture models. Results will be summarized by treatment group and visit.

8.4. Examination of Subgroups

All efficacy variables will be summarized by number of prior HER2-directed regimens in the metastatic setting, geographic region, hormone receptor status, and disease category. These analyses will be presented using forest plots and tests of interaction will be done. PFS and OS will, in addition, be summarized by de novo MBC vs. recurrent MBC.

8.5. Exploratory Biomarker Analysis

Future exploratory biomarker evaluations may include but are not limited to quantitative measures of HER2, gene expression analysis, and genomic analysis including but not limited to PIK3CA mutation status. The role of the immune response including measurement of tumor infiltrating lymphocytes (TILS) and levels of immune markers may also be assessed.

9. SAFETY EVALUATION

9.1. Overview of Safety Analysis Methods

All safety analysis will be performed for all patients in the Safety population. The following assessments will be used to evaluate the safety of neratinib in combination with capecitabine in HER2+ metastatic breast cancer (MBC) patients whose disease has progressed following two (2) prior HER2+ directed regimens in the metastatic setting:

- Adverse events (AEs)
- Medical history
- Vital sign measurements
- Physical examination findings
- Electrocardiogram (EKG)
- LVEF results from multiple-gated acquisition scan (MUGA) or echocardiogram (ECHO)
- Liver Function Tests results
- Eastern Cooperative Oncology Group (ECOG) performance status
- Laboratory assessments

All safety endpoints will be summarized by treatment group and visit when appropriate.

9.2. Extent of Exposure

Extent of exposure to each investigational product will be summarized by total dose, number of cycles, number of missed doses, number of dose delays and number of dose reductions.

9.3. Adverse Events, Serious Adverse Events, and Deaths

All AEs and SAEs will be reported until 28 days after the last dose of investigational product(s) and will be followed until resolution or until condition stabilizes. Should an Investigator be made aware of any SAE occurring any time after the reporting period, it must be promptly reported. The Patient Diary used for recording of investigational product intake will also be used by patients to document any AEs experienced during study treatment. In the case of diarrhea, it also serves to document the number of loose stools per day and use of loperamide/other antidiarrheal treatments taken.

AEs and SAEs will be coded using the most current version of MedDRA and tabulated by system organ class (SOC) and preferred term (PT). All AEs will be graded by the Investigator according to the NCI CTCAE v.4.0. Summaries will in general focus on treatment emergent adverse events (TEAEs). A TEAE is any adverse event that occurs or worsens on or after first dose of investigational product and up to 28 days after the last dose.

TEAEs and SAEs will be tabulated by SOC, PT, and treatment group. All tabulations will be sorted by descending frequency of SOC and PT in the total column unless otherwise noted. TEAEs will also be summarized according to relationship to study drug and severity. TEAEs of toxicity grade 3 or higher will also be summarized.

Similarly, treatment emergent adverse events leading to treatment discontinuation, dose modification, dose hold, and/or delay of investigational product will be tabulated by SOC, PT and treatment group.

Separate summaries and comparisons between treatment groups will be presented for treatment emergent diarrhea events leading to treatment discontinuation, dose holds, dose reductions, and delay of investigational product. Therapies used to treat diarrhea will also be summarized using descriptive statistics and compared between treatment groups.

SAEs, and TEAEs will be listed in by patient listings sorted by treatment, patient ID and study day; SAEs will be flagged. All AE listings will include study day, SOC, PT, reported term, dose, Cycle/day, AE onset date, AE resolution date, outcome, duration, relationship to drug, action taken, and severity.

Patient deaths are recorded on the End of Study CRF page. The number and incidence of death will be summarized by cause of death overall, treatment arm, and on-study status at time of death (within 28 days of last dose vs. more than 28 days after last dose). Patient death listings will include all death data available including date of death, cause of death and any AEs resulting in death.

9.4. Clinical Laboratory Evaluation

Blood samples for clinical chemistry and hematology will be collected during screening, before study drug administration, on day 1 of each treatment cycle and at the treatment discontinuation visit. Samples for pregnancy test and urinalysis will be collected at screening and before study drug administration.

Laboratory data will be summarized in tables using descriptive statistics for baseline and each cycle/visit. Descriptive statistics will be calculated on both the actual score and the change from baseline score. Additionally, clinically significant abnormalities in laboratory results will be summarized for the post-baseline cycles/visits using frequencies and percentages. Shifts in normal/abnormal status between baseline and subsequent visits will be summarized as well.

The institutional laboratory will analyze all hematology, routine blood chemistry, and urine samples collected.

9.5. Vital Signs, Physical Findings, and Other Observations Related to Safety

9.5.1. Vital Signs

Vital signs, including systolic and diastolic blood pressure, pulse, and body temperature, as well as weight will be collected during screening and on days 1, 8 and 15 of cycle 1 and on day 1 of the following cycles. Height will be collected at screening.

Summary tables will include descriptive statistics (number of patients, mean, std, median, min, and max) for baseline and each cycle/visit. Descriptive statistics will be calculated on both the actual score and the change from baseline score. Supplemental data will not be included in this summary; therefore the latest assessment in the cycle/visit window will be used for the evaluation.

A second summary will focus on the individual patient measures with the intent of identifying post-baseline abnormalities defined as clinically significant.

9.5.2. Physical Examinations

Physical examinations will be collected during screening, before study drug administration, on day 1 of each treatment cycle and at the Treatment Discontinuation Visit.

A full physical exam will be performed at Screening. Detailed/brief (system-guided) physical examinations will be done at subsequent time points to evaluate any clinically significant abnormalities, including worsening of conditions included in the patient's medical history.

Physical examination results will be listed with vital signs. Changes in weight will be summarized across time on study and the frequency of clinically significant changes will be tabulated overall and by treatment.

9.5.3. Electrocardiograms

Single standard 12 lead digital EKGs will be performed during screening, at Cycle 3 Day 1, Cycle 6 Day 1, Day 1 of every 6^{th} cycle after that, and on the treatment discontinuation visit if not done within the previous 12 weeks.

The EKG (measured after resting in a supine position for 5 minutes) will include heart rate, rhythm, and RR, PR, QRS, QT and QTc intervals. The EKG will be read and interpreted at the investigational site for patient safety monitoring, and documentation stored with the source documents.

Additional correction formulas included but considered secondary are QTcF and QTcB. These additional corrected QT intervals are defined as:

- QTcB is the length of the QT interval corrected for heart rate by Bazett's formula: $QTcB = QT / (RR)^{1/2}$
- QTcF is the length of the QT interval corrected for heart rate by Fridericia's formula: $QTcF = QT / (RR)^{1/3}$

All EKG parameters and their change from baseline will be summarized across study time points by treatment arm and overall using descriptive statistics. The frequency of abnormal EKG events will be tabulated by treatment arm and overall at each time point.

All EKG values will be listed by patient. Change in EKG findings will also be presented in a separate by-patient listing.

9.5.4. Left Ventricular Ejection Fraction (LVEF)

MUGA scan or ECHO scans to determine LVEF will be performed during screening and repeated at Cycle 3 Day 1, Cycle 6 Day 1, Day 1 of every 6th cycle after that, and at treatment discontinuation, if not done within the previous 12 weeks.

MUGA scans or ECHO scans to determine LVEF will be performed as part of Screening procedures within 2 weeks before Cycle 1/Day 1, unless results are already available from within 4 weeks before Cycle 1/Day 1.

The mean LVEF and mean LVEF change from baseline will be summarized across study time points by treatment arm and overall using descriptive statistics.

All MUGA and ECHO values will be listed by patient. Change in MUGA and ECHO findings will also be presented in a separate by-patient listing.

9.5.5. Liver Function Test Results

A listing will be generated for patients potentially meeting Hy's Law criteria (total bilirubin >= (2xULN) and AST or ALT > (3xULN) at any time during the study). Plots or tables of on-therapy peak total bilirubin vs. peak AST (or ALT) will also be generated to help identify potential Hy's Law cases.

9.5.6. ECOG Performance Status

ECOG performance status will be assessed at screening, Day 1 of every cycle, and at treatment discontinuation in accordance with the Schedule of Procedures in the study protocol. The ECOG categories are also summarized in the study protocol.

Screening ECOG performance status may be accepted as the Baseline status if the assessment was performed within 72 hours of initiation of investigational product and there are no clinically significant findings.

The ECOG status will be included in the baseline and demographic variables. The number and percentages of patients in each ECOG category will be presented by treatment and overall.

10. INTERIM ANALYSES AND DATA MONITORING

No interim PFS analysis is planned. One OS interim analysis is planned at the time of the primary PFS analysis. An Independent Data Monitoring Committee will review safety data on an ongoing basis throughout the study.

11. CHANGES TO THE ANALYSES PLANNED IN THE PROTOCOL

This section is to be filled out if there are any changes to the analyses prior to finalization of the SAP.

12. LIST OF PLANNED TABLES

Grouping	Title
Disposition	Summary of Subject Disposition
	Summary of Protocol Deviations
Demography	Summary of Demographics and Baseline Characteristics
Medical History	Summary of Medical History
Cancer History	Summary of Prior Cancer History and Therapies
	Summary of Baseline Breast Cancer Disease Burden
	Summary of Prior Anticancer Medications
	Summary of Prior Anticancer Radiotherapy
	Summary of Prior Anticancer Surgical Therapy
	Summary of Biomarkers
Concomitant Medications	Summary of Concomitant Medications
	Summary of Non-Pharmacologic Anti-Diarrheal Treatments
Study Medication	Summary of test article exposure
	Summary of test article dose reduction
	Summary of test article dose hold
Efficacy	Summary of Progression-Free Survival
	Summary of Censoring in Progression-Free Survival
	Summary of Objective Response Rate
	Summary of Duration of Response
	Summary of Censoring in Duration of Response
	Summary of Clinical Benefit Rate
	Summary of Overall Survival
	Summary of Censoring in Overall Survival
Health Outcomes	Summary of EORTC QLQ C30 BR23 Observed Results
	Summary of EORTC QLQ C30 BR23 Change from baseline
	Summary of EORTC QLQ C30 BR23 Mixed Model Between Treatment Comparison
	Summary of EQ-5D Observed Results
	Summary of EQ-5D Change from baseline

Grouping	Title
	Summary of EQ-5D Mixed Model Between Treatment Comparison
	Summary of EQ-VAS Observed Results
	Summary of EQ- VAS Change from baseline
	Summary of EQ- VAS Mixed Model Between Treatment Comparison
Safety	Frequency Table of Treatment Emergent Adverse Events by Body System and Preferred Term
	Frequency Table of Serious Adverse Events by Body System and Preferred Term
	Frequency Table of Treatment Emergent Adverse Events Leading to Study Discontinuation by Body System and Preferred Term
	Frequency Table of Treatment Emergent Adverse Events Leading to Study Drug Reduction by Body System and Preferred Term
	Frequency Table of Treatment Emergent Adverse Events Leading to Study Drug Hold by Body System and Preferred Term
	Frequency Table of Treatment Emergent Adverse Events by Body System, Preferred Term and Relationship to Study Medication
	Frequency Table of Treatment Emergent Adverse Events by Body System, Preferred Term and Severity
	Frequency Table of Treatment Emergent Adverse Events by Body System and Preferred Term (Toxicity Grade 3 or Higher)
	Summary of Death
Diarrhea	Summary of Time to First Diarrhea Treatment Emergent Adverse Event
	Diarrhea Characteristics
	Summary of Time to First Diarrhea Treatment Emergent Adverse Event by Severity
	Summary of Diarrhea Treatment Emergent Adverse Events Leading to Treatment Discontinuation
	Summary of Diarrhea Treatment Emergent Adverse Events Leading to Dose Hold
	Summary of Diarrhea Treatment Emergent Adverse Events Leading to Dose Reduction
	Summary of Anti-diarrheal Medications
Laboratory	Mean and Mean Change from Baseline in Numeric Laboratory Data: Hematology/ Clinical Chemistry/Urinalysis
	Frequency Table of Potentially Clinically Significant Laboratory Data: Hematology/ Clinical Chemistry/Urinalysis

Grouping	Title
	Shift from Baseline in Laboratory Data: Hematology/ Clinical Chemistry/Urinalysis
Vital Signs	Mean and Mean Change from Baseline in of Vital Signs Data
	Frequency Table of Potentially Clinically Significant Vital Signs Data
	Mean and Mean Change from Baseline in ECG Data
Physical Exam	Summary of Physical Examination
ECG	Frequency Table of Potentially Clinically Significant Values in QTc(F) Interval at Post First Dose
	Frequency Table of Potentially Clinically Significant Values in QTc(B) Interval at Post First Dose
	Frequency Table of Potentially Clinically Significant Values in QT Interval at Post First Dose
LVEF	Mean and Mean Change from baseline in LVEF Data

13. LIST OF PLANNED FIGURES

Grouping	Title
Efficacy	Kaplan-Meier Plot of Progression Free Survival (centrally assessed)
	Kaplan-Meier Plot of Overall Survival
	Kaplan-Meier Plot of Progression Free Survival (locally assessed)
	Kaplan-Meier Plot of Duration of Response
	Kaplan-Meier Plot of intervention for CNS disease

14. LIST OF PLANNED DATA LISTINGS

Grouping	Title
Disposition	Listing of Screen Failures
Demography	Listing of Demographics and Baseline Characteristics
Medical History	Listing of Medical History
Cancer History	Listing of Prior Cancer History
	Listing of Prior Cancer Therapies
Concomitant Medications	Listing of Concomitant Medications
	Listing of Non-Pharmacologic Anti-Diarrheal Treatments
Study Medication	Listing of Test Article Administration
	Listing of Subjects with Dose Reduction
	Listing of Subjects with Dose Hold
Efficacy	Listing of Tumor Data (centrally assessed)
	Listings of Tumor Measurements and Assessments
Health Outcomes	Listing of EORTC QLQ C30 and BR23 data
	Listing of EQ-5D data
Safety	Listing of Adverse Events
	Listing of Adverse Events with Toxicity Grade 3 or higher
	Listing of Treatment Emergent Adverse Events with Toxicity Grade 3 or higher
	Listing of Drug Related Adverse Events
	Listing of Serious Adverse Events
	Listing of Adverse Events Leading to Treatment Discontinuation
	Listing of Adverse Events Leading to Dose Reduction
	Listing of Adverse Events Leading to Dose Hold
	Listing of Drug Related Adverse Events
	Listing of Deaths
	Listing of potential Hy's law cases
Diarrhea	Listing of Treatment Emergent Diarrhea by Severity
	Listing of Treatment Emergent Diarrhea Grade 3 or Higher
	Listing of Anti-diarrheal Medications
Laboratory	Listing of Laboratory Data

Grouping	Title
Vital Signs	Listing of Vital Signs Data
	Frequency Table of Potentially Clinically Significant Vital Signs Data
Physical Exam	Listing of Physical Examination Data
ECG	Listing of ECG Data
LVEF	Listing of LVEF Data

15. REFERENCES

De Mets DL and Gordon Lan KK. Interim analysis: the alpha spending function approach. Stat Med. 1994;13(13-14):1341-56.

Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 1988; 16:1140-54.

Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43(3):203-220.

Ye Y, Li A, Liu L, and Yao B. A group sequential Holm procedure with multiple primary endpoints, Stat Med 2013 Mar 30;32(7):1112-24.

SUPLEMENTAL STATISTICAL ANALYSIS PLAN FOR THE NALA (1301) STUDY

STUDY DRUG

Neratinib

VERSION 1.0
DATE OF PLAN

14-NOV-2018

SPONSOR

Puma Biotechnology, Inc.

10880 Wilshire Blvd, Suite 2150

Los Angeles, CA 90024 USA

Phone: +1 424.248.6500
Fax: +1 424.248.6501

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Term
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BR-23	European Organization for Research and Treatment-QOL questionnaire breast cancer specific module
CNS	Central Nervous System
EOT	End of treatment
EQ-5D-5L	EuroQol Group health status questionnaire
HOA	Health Outcomes Assessment
KM	Kaplan-Meier
OS	Overall Survival
PFS	Progression-free Survival
QLQ-C30	European Organization for Research and Treatment-QOL questionnaire
QOL	Quality of Life
SAP	Statistical Analysis Plan
TTD	Time to Deterioration
UK	United Kingdom

1. INTRODUCTION

This supplemental Statistical Analysis Plan (SAP) is intended to further describe the efficacy analysis for the secondary endpoints of Health Outcomes Assessments (HOA) and Central Nervous System (CNS) as well as some additional sensitivity analyses.

For the purposes of this supplemental SAP we will use the term Quality of Life (QoL) to cover all the analysis relating to the Health Outcomes Assessments (HOA).

2. ANALYSIS OF QUALITY OF LIFE DATA

This section expands on the QoL analysis described in sections 8.3.3.1-8.3.3.2 of the Statistical Analysis Plan (SAP) version 2.1.

In summary, there were three Quality of Life (QoL) instruments used in the NALA study: QLQ-C30, BR-23, and EQ-5D-5L. There are 26 scales in these instruments and they are presented in Table 1 (QLQ-C30: 16 scales, BR-23: 8 scales, VAS and EQ-5D-5L health index score). The HOA instruments are administered every 2 cycles (6 weeks) until the end of study drug administration.

Table 1: Quality of Life Instruments and Scales

	Scale	Number of items	Score Range	Version 3.0 Item Numbers	Function or Symptom Scales	
QLQ-C30 Summary Score	C30Sum	13	0-100	1-28, 30	F	QLQ-C30
Global Health status/QoL						
Global health status/QoL	QL2	2	0-100	29, 30	F	QLQ-C30
Functional Scales						
Physical functioning	PF2	5	0-100	1 to 5	F	QLQ-C30
Role Functioning	RF2	2	0-100	6,7	F	QLQ-C30
Emotional Functioning	EF	4	0-100	21 to 24	F	QLQ-C30
Cognitive Functioning	CF	2	0-100	20, 25	F	QLQ-C30
Social Functioning	SF	2	0-100	26, 27	F	QLQ-C30
Symptom Scales						
Fatigue	FA	3	0-100	10, 12, 18	S	QLQ-C30
Nausea and vomiting	NV	2	0-100	14, 15	S	QLQ-C30
Pain	PA	2	0-100	9, 19	S	QLQ-C30
Dyspnoea	DY	1	0-100	8	S	QLQ-C30
Insomnia	SL	1	0-100	11	S	QLQ-C30
Appetite loss	AP	1	0-100	13	S	QLQ-C30
Constipation	CO	1	0-100	16	S	QLQ-C30
Diarrhea	DI	1	0-100	17	S	QLQ-C30
Financial difficulties	FI	1	0-100	28	S	QLQ-C30

Table 1: Quality of Life Instruments and Scales (continued)

	Scale	Number of items	Score Range	Version 3.0 Item Numbers	Function or Symptom Scales	
Functional Scales			0-100			
Body Image	BRBI	4	0-100	9-12	F	BR23
Sexual Functioning	BRSEF	2	0-100	14, 15	S	BR23
Sexual Enjoyment	BRSEE	1	0-100	16	S	BR23
Future Perspective	BRFU	1	0-100	13	F	BR23
Symptom Scales/Items			0-100			
Systemic therapy side effects	BRST	7	0-100	1-4, 6, 7, 8	S	BR23
Breast Symptoms	BRBS	4	0-100	20-23	S	BR23
Arm Symptoms	BRAS	3	0-100	17, 18, 19	S	BR23
Upset by hair loss	BRHL	1	0-100	5	S	BR23
EQ-5D-5L Health State Summary Score	VAS		-0.595 – 1.000		F	EQ-5D-5L
EQ-5D-5L Index Score	IndexS		0-100		F	EQ-5D-5L

2.1.1. QoL populations

In the 1301 SAP version 2.1 the QoL population was defined as a single population across all the HOA instruments. After the approval of the SAP it was determined that it would be best to define a separate population for each of the subscales. A subject will be included in the population for a particular subscale if they are in the safety population and they had a baseline assessment for that subscale as well as at least one post baseline assessment (up to last dose day +28 days) for the specific subscale. In the analysis dataset there will be a flag for each subscale population.

2.2. QLQ-C30 and BR-23 Scoring

The scoring for QLQ-C30 and BR-23 follows the EORTC QLQ-C30 scoring manual in following link: https://www.eortc.be/qol/files/SCManualQLQ-C30.pdf

2.2.1. EQ-5D-5L Scoring

The EQ-5D-5L consists of 6 items: 5 dimensions (mobility, self-care, usual activity, pain/discomfort, and anxiety/depression) and a health state score. Each of the 5 dimensions has 5 levels. The combination of 1 level from each of the 5 dimensions (e.g. 12234) will be used to obtain the health index score.

To obtain EQ-5D-5L Index score, it is necessary to map to a value set from a particular country that has been well established and validated. The EQ-5D-5L cross-walk value set for United Kingdom (UK) will be used in the derivation of health index scores. UK value sets has been used by PUMA studies 3004 and 6201 in alignment with wide utilization of EQ-5D results in UK.

Derivation of the EQ-5D-5L index scores will use the "EQ-5D-5L Crosswalk value sets" link at the following location:

https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/valuation/crosswalk-index-value-calculator/

2.3. Assessments on the same date

Multiple assessments on the same date will be averaged so only one score will be used for any subscale for the given date.

2.4. Visit Time Window

Each QoL assessment is mapped to a scheduled visit. The defined visit windows are non-overlapping and cover all possible dates from Baseline through the date of last dose of study drug administration plus 28 days.

2.4.1. Visit windows:

- The Baseline assessment will be the closest assessment on or before the date of first dose of study drug.
- The time window for C3D1 will be first dose date+1 through first dose date +62 days.
- Except Baseline and C3D1, the time window for the rest of visits will be defined as:
  - Start of window: study day of (1+21*(# of cycles)-21))
  - End of visit window: study day of (1+21*(# of cycles) +20)).
  - The target visit day for a scheduled visit is: (1+21*(# of cycles)).

The visit time window mapping up to cycle 19 can be found below in reference to Day 1 which is the first dose date (Day 1).

Cycle	Target Day	Start Day	End Day
Baseline	Day 1	< Day 1	Day 1
Cycle 3 Day 1	Day 43	Day 2	Day 63
Cycle 5 Day 1	Day 85	Day 64	Day 105
Cycle 7 Day 1	Day 127	Day 106	Day 147
Cycle 9 Day 1	Day 169	Day 148	Day 189
Cycle 11 Day 1	Day 211	Day 190	Day 231
Cycle 13 Day 1	Day 253	Day 232	Day 273
Cycle 15 Day 1	Day 295	Day 274	Day 315

Cycle	Target Day	Start Day	End Day
Cycle 17 Day 1	Day 337	Day 316	Day 357
Cycle 19 Day 1	Day 379	Day 358	Day 399

2.4.2. Multiple Assessments within a visit window

If a subject has multiple assessments on different dates within a visit window the assessment closest to the target visit day will be used for descriptive analysis/plotting, mixed model and response analysis by visits where a single value is associated with a visit.

2.5. Statistical Analysis

2.5.1. Descriptive Analysis

Observed scores and change from baseline will be summarized by treatment group and visit for all 26 subscales. Results will be summarized in a table and figure for each subscale. The tables will include all time points available in the data and the plots are to include time points through cycle 19 (approximately 12.5 months).

2.5.2. Mixed Model Analysis

A mixed model analysis with an outcome of change from baseline and the covariates including baseline score, treatment arm, visit (categorical), treatment arm by time interaction, prior HER2-directed regimens in the metastatic setting (2 or 3+), hormone receptor status (positive or negative), and disease location (visceral and non-visceral) will be used to evaluate differences between treatment arms. The analysis will use the F test from the repeated measures mixed model. Three randomization factors are included in the model to be consistent with the primary analysis as documented in a note to file (dated 8-SEP-2018) to the main SAP.

A reduced mixed model will be used to estimate the change over time by treatment group. These results will be summarized in a table and figure for each subscale. The reduced mixed model will only include baseline score, treatment arm, time (categorical), and treatment arm by time interaction.

2.5.3. Response Analysis

A response analysis will be used to evaluate the score change over time with a minimum clinical important threshold value. At each post-baseline assessment, a subject's score will be classified as having a response of improved, stable, or worsened. A Mantel-Haenszel Chi Square test to assess any differences between treatment arms will be done at each visit and the results will be summarized in a table.

We will conduct the response analysis for the following 7 subscales:

- Global Health
- C30 summary score,
- Physical functioning
- Fatigue

- Constipation
- Diarrhea
- Systemic Therapy Side Effects scale

For the functional scales and the C30 summary score, an improvement is defined as an increase in 10 points or more, worsening is defined as a decrease of 10 points or more, and stable is defined as neither improved nor worsened.

For the symptom scales, an improvement is defined as a decrease of 10 points or more, worsening is defined as an increase of 10 points or more, and stable is defined as neither improved nor worsened. The results will be summarized in a table for each of the seven scales.

2.5.4. Time To Deterioration (TTD) Analysis

For TTD analysis, all records (including multiple assessments within a visit window post baseline assessment) sorted in chronological order will be used to identify the first event. This analysis will be done for the 7 subscales listed in Section 2.5.3

Time to deterioration is defined as the time from Baseline to:

1. Function scales: to the first assessment date where a 10 point or more decrease was observed for the scale of interest
2. Symptom scales: to the first assessment date where a 10 point or more increase was observed for the scale of interest.

If subject's score change did not reach the deterioration threshold value they will be censored at their last HOA assessment. If a subject dies (occurred on or before last dose date plus 28 days) prior to a documented decline in HOA assessment, the subject will be considered to have the deterioration event on the death date unless the date of death occurs after 2 missed HOA assessments (12 weeks + 3 days), in that case, the subject will be censored at last QOL assessment prior to death.

The log-rank test from a Kaplan-Meier method will be used to assess treatment differences. In addition a Cox Proportional Hazards model including treatment arm, prior HER2-directed regimens in the metastatic setting (2 or 3+), hormone receptor status (positive or negative), and disease location (visceral and non-visceral) will be used to estimate the hazard ratio. The use of three of the four randomization strata is documented in a Note to file dated 6-SEP-2018.

2.5.5. Area Under Curve (AUC) analysis

Time adjusted AUC will be derived for each of the subscales listed in Section 2.5.3

The time adjusted AUC is calculated using the trapezoidal rule between adjacent time points using the actual assessment study day for each subject by subscale. This method sums up the areas cumulated through multiple consecutive time point assessments from baseline assessment day to last assessment day. The area between two time points is calculated by taking the average subscale score between the 2 time points and multiplied by the duration between the two time points.

All assessments post baseline will be utilized except when there are multiple assessments on the same day, in which case, only the derived average subscale for the day in question will be used.

Time-adjusted area under the curve = Area under the Curve/(duration from baseline assessment to last assessment in days)

Comparison of AUC by treatment group will be summarized in a table using an ANCOVA model to test for treatment difference of AUC. The model will include the baseline subscale value and treatment arm.

2.5.6. Completion rate

A completion analysis will be done to evaluate the magnitude of missing assessments. The completion analysis will focus on the following scales.

1. QLQ-C30 – the C30 summary subscale (required non-missing subscale scores from all except financial difficulty, the most conservative scale from C30)
2. BR-23 - the systemic therapy sides effects subscale
3. EQ5D-5L - either the VAS or Index Score

Completion is defined as:

- Overall: The actual number of assessments (baseline and all post baseline assessments) / expected number of assessments (from baseline to last assessment thru last dose date+28 days)*100%.
- Calculation for expected number of assessments will be based on treatment duration and corresponding cycle of last dose day. If last dose day falls in a day in cycle Y which is the completion of an even cycle , the expected number of HOA assessments will be the integer part of (Y/2) + 1; or else If last dose day falls in a cycle Y and Y/2 does not yield a whole number then , the expected number of HOA assessments will be the integer portion of (Y/2) + 2.

  For example, a subject with the last dose in cycle 4 would be expected to have 3 assessments (Baseline, cycle 3 and EOT). If there are multiple assessments on the same day only one assessment will be counted for that day. Lastly, if the completion is greater than 100%, it will be truncated at 100%.
- Completion at each scheduled visit: The number of subjects that completed the scale flagged for the scheduled visit divided by the number of subjects under treatment in the corresponding cycle.

The overall completion rate and completion rate by visit will be summarized by treatment group for each relevant scale.

2.5.7. Pattern Mixture Model Analysis

After reviewing findings from the descriptive statistics, the overall completion rate, and the scheduled visit specific completion rate analysis if there exists a pattern of missing assessments additional sensitivity analysis like pattern mixture model analysis may be added in an ad hoc fashion.

3. ANALYSIS OF THE CNS ENDPOINTS

Additional details on the CNS endpoints are provided in this section.

3.1. Secondary CNS endpoint

The secondary CNS endpoint is time to intervention for symptomatic CNS disease. This is defined as the time from the date of randomization to the first date of an intervention for symptomatic metastatic CNS disease. Interventions include medications to control signs or symptoms caused by the CNS metastases (ie, high dose steroids in case of cranial hypertension), CNS radiotherapy, CNS surgery and cancer treatment specific for CNS metastases.

3.2. Progressive CNS Endpoint

The alternative CNS endpoint is time to progressive CNS lesions. This is defined as the time interval from the date of randomization until the first date on which a scheduled or unscheduled tumor assessment documents the appearance of newly diagnosed or progressive CNS lesions, per RECIST v1.1. For subjects that have CNS lesions as target lesions, the sum of the longest diameter (SLD) of CNS lesions only will be calculated at each tumor assessment and if the change from the nadir is greater than 20% this will be considered progressive CNS disease. If the subject has nontarget CNS lesions progressive CNS disease will be considered if any of the nontarget CNS lesions have a status of unequivocal progression. If the subject had a newly diagnosed CNS lesion post-baseline this will be considered progression as well.

3.3. Statistical Analysis

3.3.1. Sensitivity analysis for the secondary CNS endpoint

1. The Kaplan Meier method, stratified log-rank test, and Cox proportional hazard model will be performed without considering competing events. A Kaplan-Meier curve with the results of the Cox model will summarize this analysis.
2. An analysis only looking at time to first dose of radiation for symptomatic CNS will be performed. Subjects that do not have a radiation intervention for symptomatic metastatic CNS and do not die will be censored at the last date known alive on or prior to the data cutoff. Deaths or other treatments for symptomatic CNS are treated as competing events.
3. A subgroup analysis only including subjects with a prior or current history of CNS metastases at the time of randomization will be done for the secondary CNS endpoint .A prior history will be defined as location of disease in the brain at screening, prior radiation of the brain, and prior surgery in the brain.

3.3.2. Time to progressive CNS disease

The time to progressive central nervous system (CNS) lesions will be analyzed using the centrally reviewed tumor assessments as well as the investigator tumor assessments.

Time to progressive CNS lesions is defined as the time interval from the date of randomization until the first date on which a scheduled or unscheduled visit documents the appearance of newly

diagnosed or progressive CNS lesions, per RECIST v1.1. Table 4 lists the CNS locations for the derivation of CNS progression based on the central or local tumor assessment data. All progressive CNS events will be counted regardless of their timing with regards to progression at other sites. Competing events include progression at other sites and deaths prior to the observance of progressive CNS. If there is no evidence of CNS progression or a competing event, the subject will be censored at the last assessable tumor evaluation on study or prior to new anti-cancer therapy, if applicable. Gray's test (Gray, 1988) stratified for factors used in the randomization will assess the equality of cumulative incidence functions between the treatment arms. The cause specific hazard ratio and p-value will also be performed. Additionally, the Kaplan Meier method, stratified log-rank test, and stratified Cox proportional hazard model will be performed without considering competing events. For this analysis subjects will be censored at the time of PD at a site other than CNS or at their last viable tumor assessment for subjects that did not have a PD. For these CNS endpoints the same censoring rules as those specific in the main SAP for the PFS analysis will be applied. Events, CNS events or any competing events, after the start of a new anti-cancer therapy or after 2 missed assessments will be censored and not considered as events.

4. SENSITIVITY ANALYSIS FOR OVERALL RESPONSE RATE

The SAP indicates that only subjects with measurable disease at baseline will be used in the calculation of Overall Response Rate. As a sensitivity analysis, all subjects will be included in the calculation. For subjects that only have non-target lesions at baseline, they can only be considered a responder if they achieve CR in all non-target lesions.

5. PROGNOSTIC VARIABLES OF INTEREST

The following prognostic variables will be investigated for the primary endpoints of PFS and OS.

1. Previous HER2 directed therapies
    a. Trastuzumab only
    b. TDM-1 only
    c. Trastuzumab and Pertuzumab
    d. Trastuzumab and T-DM1
    e. Trastuzumab and Pertuzumab and T-DM1
2. De Novo and non De Novo disease at diagnosis

    De novo disease at diagnosis is defined as having M stage of M1 or if the date of initial diagnosis is equal to or within 45 days of the date of first presentation with distant metastatic disease.

6. ADDITIONAL EXPLORATORY ANALYSIS

The following exploratory analyses for PFS will be performed.

6.1. Time to RECIST defined PFS or clinical progression

For this analysis all the definitions stated in the main SAP apply but the definition of PD is expanded to include subjects that ended treatment for clinical progression and did not experience PD per RECIST.

6.2. Time to treatment failure

Time to treatment failure is defined as the time from randomization to treatment discontinuation for any reason. For those subjects that are still on treatment at the time of the analysis, they will be censored on the data cut-off date.

Table 2: List of Planned Tables

Grouping	Title
HOA analysis	
Descriptive HOA	Summary of scale scores and change from baseline at each cycle (26 tables)
HOA Mixed model	Summary of model results for each subscale (26 tables)
HOA reduced mixed model	Summary of estimated means at each cycle for each subscale (26 tables)
Response analysis	Summary of response by subscale (7 tables)
AUC Analysis	Summary of ANCOVA model results (1 table)
Completion analysis	Summary of overall completion (1 table)
	Summary of completion by visit (1 table)
CNS Analysis	
Time to intervention for symptomatic CNS	Summary of time to intervention for symptomatic CNS
CNS disease progression	Summary of CNS Progression-Free Survival – Central Assessment
	Summary of CNS Progression-Free Survival – Local Assessment
CNS disease progression – no competing events	Summary of CNS Progression-Free Survival – Central Assessment
	Summary of CNS Progression-Free Survival – Local Assessment
PFS sensitivity analysis	Summary of Progression-free survival including EOT for clinical progression.
Sensitivity for ORR	Summary of ORR and CBR (ITT Population)
Time to treatment failure	Summary of time to Treatment Failure

Table 3: List of Planned Figures

Grouping	Title
QOL analysis	
Descriptive QOL	Figure of scale scores at each cycle (26 figures)
	Figure of change from baseline in each scale (26 figures)
Estimated means from mixed model	Mixed Model Results by Cycle (Change from Baseline) (26 figures)
Time to deterioration	KM curve by treatment group (7 figures)
CNS Analysis	
Time to intervention for symptomatic CNS	Cumulative hazard curves
	KM curve for no competing events
CNS disease progression – with competing risks	Cumulative hazard curve for CNS Progression-Free Survival – Central Assessment
	Cumulative hazard curve for CNS Progression-Free Survival – Local Assessment
CNS disease progression – no competing events	KM curve for CNS Progression-Free Survival – Central Assessment
	KM curve of CNS Progression-Free Survival – Local Assessment
Other	
KM of time to treatment discontinuation	KM curve of time to treatment discontinuation
Sensitivity for PFS	KM curve of time to PFS including clinical progression
Forest Plot for prognostic factors – PFS	Forest plot including prior treatments and DeNovo Disease for Central PFS
Forest Plot for prognostic factors – PFS	Forest plot including prior treatments and DeNovo Disease for Local PFS
Forest Plot for prognostic factors – OS	Forest plot including prior treatments and DeNovo Disease for OS

Table 4: List of CNS locations

	Bioclinica (Central review)	**INFORM (Local review)**
CNS Locations	Basal Ganglia Brainstem Caudate Cerebellum CNS (Brain) Corpus Callosum Dural Epidural Falx Fourth Ventricle Frontal Lobe Frontoparietal Hippocampus Lateral Ventricle Leptomeningeal Occipital Lobe Parietal Lobe Parietoccipital Spinal Cord Subdural Temporal Lobe Temporaloccipital Temporalparietal Thalamus Third Ventricle	Brain CNS Retina