AstraZeneca

Statistical Analysis Plan

Study Code	PT009003
Version	1.0
Date	09 February 2018

A Randomized, Double-Blind, Parallel Group, Multi-Center Study to Assess the Efficacy and Safety of PT009 compared to PT005 in Subjects With Moderate to Very Severe COPD

STATISTICAL ANALYSIS PLAN FOR STUDY PT009003

Protocol Number:	PT009003-03
Investigational Drug and Drug Number:	BFF MDI; PT009 FF MDI; PT005
Indication:	COPD
Dosage Form/Dose:	BFF MDI 320/9.6 µg BID BFF MDI 160/9.6 µg BID FF MDI 9.6 µg BID

PT009003 Protocol Title: A Randomized, Double-Blind, Parallel Group, Multi-Center Study to Assess the Efficacy and Safety of PT009 Compared to PT005 in Subjects with Moderate to Very Severe COPD

Date of Issue: 09 February 2018

Version: Version 1.0

Signed Agreement on Statistical Analysis Plan

FINAL SIGN-OFF SIGNATURES

Primary Biostatistician and Peer Reviewer:	[redacted]
Author:	
Approved by:	
Approved by:	
Approved by:	
Approved by:	

Change Log			
Version No.	**Effective Date**	**Reason for the Change / Revision**	**Supersedes**

TABLE OF CONTENTS

1. INTRODUCTION 12

2. STUDY OBJECTIVES AND ENDPOINTS 12

2.1 STUDY OBJECTIVES 12
2.1.1 Primary Objective 12
2.1.2 Secondary Objectives 12
2.1.3 Safety Objectives 12
2.1.4 Healthcare Resource Utilization (HCRU) Objective 12
2.2 STUDY ENDPOINTS 12
2.2.1 Primary Efficacy Endpoints 13
2.2.2 Secondary Efficacy Endpoints 13
2.2.3 Other Efficacy Endpoints 13
2.2.4 Safety Endpoints 15
2.2.5 Health Care Resource Utilization Endpoints 15

3. STUDY DESIGN AND ANALYTICAL CONSIDERATIONS 16

3.1 STUDY DESIGN 16
3.1.1 Overall Study Design and Plan 16
3.1.2 Prior, Concomitant, Post-Treatment, Prohibited Medications, and Other Restrictions 18
3.2 HYPOTHESIS TESTING 18
3.3 INTERIM ANALYSIS 19
3.4 SAMPLE SIZE 19

4. DATA AND ANALYTICAL QUALITY ASSURANCE 21

5. ANALYSIS POPULATIONS 21

5.1 POPULATION DEFINITIONS 21
5.1.1 Intent-to-Treat (ITT) Population 21
5.1.2 Modified Intent-to-Treat (mITT) Population 21
5.1.3 Rescue Ventolin User Population 22
5.1.4 Safety Population 22
5.2 POPULATIONS FOR PRIMARY AND SENSITIVITY ANALYSES 22

6. STATISTICAL ANALYSIS 22

6.1 DATA HANDLING RULES AND DEFINITIONS, INCLUDING HANDLING OF MISSING DATA 23
6.2 SUBJECT DISPOSITION AND ANALYSIS POPULATIONS 24
6.3 DEMOGRAPHIC AND BASELINE CHARACTERISTICS AND EXTENT OF EXPOSURE 25
6.3.1 Demography, Baseline Characteristics, and CAT 25
6.3.2 COPD History, Screening/Baseline Spirometry, and Reversibility 26
6.3.3 Medical and Surgical History at Screening, Reproductive Status and Pregnancy Testing 26
6.3.4 Prior, Concomitant, and Post-Treatment Medications/Treatments 27
6.3.5 Extent of Exposure to Study Medication and Compliance 28
6.4 EFFICACY ANALYSES 28
6.4.1 Estimands 28
6.4.2 Baselines and Baseline Covariates for Analysis 29
6.4.3 Impact of Variable Length of the Study on the Analyses 30
6.4.4 Primary Efficacy Analysis 31
6.4.4.1 Change from Baseline in Morning Pre-Dose Trough FEV_1 31
6.4.4.2 Assumptions Checks and Removal of Outliers in Sensitivity Analyses 32
6.4.4.3 Sensitivity Analyses for Missing Data 32
6.4.4.4 Cumulative Responder Analysis 34
6.4.5 Secondary Efficacy Analyses 35

6.4.5.1 Time to First Moderate or Severe COPD Exacerbation....35
6.4.5.2 Time to First Clinically Important Deterioration....38
6.4.5.3 Transition Dyspnea Index....38
6.4.5.4 St. George's Respiratory Questionnaire....40
6.4.5.5 Rescue Ventolin HFA Use....41
6.4.5.6 EXACT Total Score....43
6.4.6 Other Efficacy Analyses....44
6.4.6.1 Other Spirometry Endpoints....44
6.4.6.2 Rate of COPD Exacerbations....44
6.4.6.3 Time to First COPD Exacerbation of any Severity....49
6.4.6.4 Time to First Severe COPD Exacerbation....49
6.4.6.5 Time to First COPD Exacerbation Treated with Systemic Steroids....50
6.4.6.6 Time to First COPD Exacerbation Treated with Antibiotics....50
6.4.6.7 Time to Treatment Failure....50
6.4.6.8 Time to Death: All Causes, Respiratory....50
6.4.6.9 Time to First Sustained Clinically Important Deterioration....50
6.4.6.10 European Quality-of-Life-5 Dimension-5 Level Questionnaire (EQ-5D-5L)....51
6.4.7 Subgroup Analyses....52
6.4.8 Control of Type I Error....53
6.4.8.1 US Approach....53
6.4.8.2 Ex-US Approach....55
6.4.9 Correlation Analysis....57
6.5 SAFETY ANALYSIS....57
6.5.1 Adverse Events....58
6.5.1.1 Adverse Events of Special Interest....60
6.5.1.2 Major Adverse Cardiovascular Event (MACE) Events Determined by Clinical Endpoint Committee....62
6.5.1.3 Pneumonia Events Determined by Clinical Endpoint Committee....62
6.5.1.4 Cause of Death Determined by Clinical Endpoint Committees....63
6.5.1.5 Paradoxical Bronchospasms....63
6.5.2 Clinical Laboratory Measurements....63
6.5.3 Vital Signs....67
6.5.4 12-Lead Electrocardiogram Measurements....69
6.5.5 Healthcare Resource Utilization....70
6.5.6 Physical Examination....72

7. CHANGES FROM METHODS PLANNED IN THE PROTOCOL 72

8. STATISTICAL SOFTWARE 73

9. REFERENCES 73

APPENDIX 1: DATA HANDLING RULES 75

APPENDIX 2: ANALYSIS DATASET SPECIFICATIONS 83

APPENDIX 3: SAS CODE FOR STATISTICAL ANALYSES 83

APPENDIX 4: CTCAE LABORATORY TEST CRITERIA FOR SHIFT TABLES AND CENTRAL LABORATORY REFERENCE RANGES FOR USE IN FLAGGING ABNORMAL VALUES 83

APPENDIX 5: STANDARD MEDDRA QUERIES (SMQ) 83

APPENDIX 6: STATISTICAL DETAILS 83

APPENDIX 7: TABLE OF CONTENTS END-OF-TEXT TLFS 84

APPENDIX 8: EQ-5D CROSSWALK SPSS AND SAS CODE FOR STATISTICAL ANALYSES OF INDEX SCORE 115

List of Tables

Table 1 Planned Sample Size....17
Table 2 Power Estimates for Morning Pre-Dose Trough FEV_1 19
Table 3 Power Estimates for Incidence of Moderate or Severe COPD Exacerbations....20
Table 4 Sensitivity Analyses for Morning Pre-Dose Trough FEV_1 33
Table 5 Sensitivity Analyses for Time to First Moderate or Severe COPD Exacerbations 36
Table 6 Adverse Events of Special Interest 60
Table 7 Lab Parameters....64
Table 8 Potentially Clinically Significant (PCS) Laboratory Parameter Criteria 66
Table 9 Potentially Clinically Significant Criteria for Systolic and Diastolic Blood Pressure Parameters 68
Table 10 Potentially Clinically Significant Criteria for Heart Rate Parameters 68
Table 11 Analysis Study Time Windows for Vital Signs Assessments....69
Table 12 Criteria for PCS ECG Values....70

List of Figures

Figure 1 Study Design....18
Figure 2 Overarching Intervals of Moderate-or-Severe (QMS) and Severe (QS) COPD Exacerbations....47
Figure 3 Overarching Intervals (I) of Mild-Moderate-or-Severe COPD Exacerbation Events Based on e-Diary Symptom Data....48
Figure 4 Overarching Intervals (QQ) of Mild-Moderate-or-Severe COPD Exacerbation Events Incorporating Both CRF Data and e-Diary Symptom Data 48
Figure 5 Group 1: Type I Error Control for the Analyses of the Primary Endpoints (US Approach) 54
Figure 6 Group 2: The Analysis of the Secondary Endpoints (US Approach) 55
Figure 7 Group 1: Type I Error Control for the Analyses of the Primary Endpoints (Ex-US Approach) 56
Figure 8 Group 2: The Analysis of the Secondary Endpoints (Ex-US Approach)....57

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

AE	Adverse Event
AESI	Adverse Event of Special Interest
ALT	Alanine Aminotransferase
ANCOVA	Analysis of Covariance
AR(1)	Autoregressive Order 1
AST	Aspartate Aminotransferase
ATS	American Thoracic Society
BDI	Baseline Dyspnea Index
BFF MDI	Budesonide and Formoterol Fumarate Metered Dose Inhaler
BID	*Bis In Die*, Twice Daily
bpm	Beats Per Minute
BMI	Body Mass Index
CAT	Chronic Obstructive Pulmonary Disease Assessment Test
CCU	Coronary Care Unit
CCV	Cardio- and Cerebrovascular
CD	Compact Disc
CI	Confidence Interval
CID	Clinically Important Deterioration
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration Equation (according to National Kidney Disease Education Program)
COPD	Chronic Obstructive Pulmonary Disease
CRF	Case Report Form
CSR	Clinical Study Report
CTCAE	Common Terminology Criteria for Adverse Events
DMC	Data Monitoring Committee
E-RS	Evaluating Respiratory Symptoms
ECG	Electrocardiogram

eCRF	Electronic Case Report Form
eDiary	Electronic Diary
e.g.	*Exempli Gratia*, For Example
eGFR	Estimated glomerular filtration rate
EOS	Eosinophil
EOT	End of Treatment
ERS	European Respiratory Society
EQ-5D	EuroQol 5 Dimensions Questionnaire
EQ-5D-5L	EuroQol 5 Dimensions Questionnaire 5-level
ER	Emergency Room
EXACT	Exacerbations of Chronic Pulmonary Disease Tool – Patient Reported Outcomes
FEV_1	Forced Expiratory Volume in 1 Second
FEF_{25-75}	Forced Expiratory Flow between 25% and 75% of FVC
FF	Formoterol Fumarate
FVC	Forced Vital Capacity
H_0	Null Hypothesis
H_1	Alternative Hypothesis
hCG	Human Chorionic Gonadotropin
HCRU	Health Care Resource Utilization
HFA	Hydrofluoroalkane
HLGT	High Level Group Term
HLT	High Level Term
ICF	Informed Consent Form
ICS	Inhaled Corticosteroid
ICU	Intensive Care Unit
ID	Identification
i.e.	*Id Est,* That Is

ITT	Intent-to-Treat
L	Liter
LABA	Long-Acting β_2-Agonist
LAMA	Long-Acting Muscarinic Antagonist
MACE	Major Adverse Cardiovascular Event
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MCID	Minimal Clinically Important Difference
MDI	Metered Dose Inhaler
MedDRA	Medical Dictionary for Regulatory Activities
MI	Multiple Imputation
µg	Microgram
mITT	Modified Intent-to-Treat
mL	Milliliter
mm	Millimeter
mmHg	Millimeter of Mercury
MNAR	Missing Not at Random
msec (ms)	Millisecond
NHANES III	Third National Health and Nutrition Examination Survey
OTC	Over-the-Counter
PCS	Potentially Clinically Significant
PEFR	Peak Expiratory Flow Rate
PIN	Personal Identification Number
PFT	Pulmonary Function Test
PMM	Pattern-Mixture Model
PT	Preferred Term
PT005	Formoterol Fumarate Inhalation Aerosol

PT009	Budesonide and Formoterol Fumarate Inhalation Aerosol
QoL	Quality of life
QTcF	QT Corrected Using Fridericia's Formula
RM	Repeated Measures
ROM	Read-Only Memory
RVU	Rescue Ventolin User
SABA	Short-Acting β_2-Agonist
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Standard Deviation
SGRQ	St. George's Respiratory Questionnaire
SMQ	Standard MedDRA Query
TC	Telephone Call
TDI	Transition Dyspnea Index
TEAE	Treatment-Emergent Adverse Event
ULN	Upper Limit of Normal
US	United States
VAS	Visual Analog Scale

Trademark Information

Trademarks Not Owned By Pearl

SAS® software

Ventolin®

1. INTRODUCTION

This Statistical Analysis Plan (SAP) outlines the statistical methods for the display, summary and analysis of data to be performed at the end of Pearl Therapeutics, Inc. (Pearl) Study PT009003. The SAP should be read in conjunction with the study protocol. This version of the SAP has been developed using the PT009003-03 Amended Protocol (Version 4.0 dated 08 January 2018) and the PT009003 case report form (CRF) (Revision 03 dated 04 May 2017).

2. STUDY OBJECTIVES AND ENDPOINTS

2.1 Study Objectives

The overall objective is to assess the efficacy and safety of treatment with BFF MDI 320/9.6 µg (budesonide and formoterol fumarate metered dose inhaler), BFF MDI 160/9.6 µg, and FF MDI 9.6 µg (formoterol fumarate metered dose inhaler) administered twice daily (BID) over a minimum of 12 weeks and a maximum of 52 weeks in subjects with moderate to very severe chronic obstructive pulmonary disease (COPD).

2.1.1 Primary Objective

- To assess the effects of BFF MDI relative to FF MDI on lung function

2.1.2 Secondary Objectives

- To assess the effects of BFF MDI relative to FF MDI on COPD exacerbations
- To assess the effects of BFF MDI relative to FF MDI on symptoms of COPD
- To assess the effects of BFF MDI relative to FF MDI on quality of life (QoL)

2.1.3 Safety Objectives

- To assess the safety of BFF MDI and FF MDI

2.1.4 Healthcare Resource Utilization (HCRU) Objective

- To assess overall and COPD-specific HCRU of BFF MDI and FF MDI

2.2 Study Endpoints

The primary endpoints, treatment comparisons of interest, and analysis timeframes may differ by country or region due to local regulatory agency requirements. The Sponsor has defined two different registration approaches in this study. The registration approaches will be called: (1) United States (US) and (2) Ex-US. The US approach is for countries or regions such as the United States where the primary and secondary endpoints are generally evaluated at a point in time. The Ex-US approach is for registration purposes in countries or regions such as Europe where the primary and secondary endpoints are generally evaluated over a period of time. The multiplicity controls for the primary and secondary efficacy analyses are delineated in section 6.4.8.

Note that spirometry endpoints are generally evaluated relative to baseline; thus, change from baseline is always implied for these endpoints.

2.2.1 Primary Efficacy Endpoints

US APPROACH

- Morning pre-dose trough Forced Expiratory Volume in 1 Second (FEV_1) at Week 12

Ex-US APPROACH

- Morning pre-dose trough FEV_1 over 24 weeks

2.2.2 Secondary Efficacy Endpoints

US APPROACH

- Time to first moderate or severe COPD exacerbation
- Change from baseline in average daily rescue Ventolin Hydrofluoroalkane (HFA) use over 12 weeks
- Percentage of subjects achieving a Minimal Clinically Important Difference (MCID) of 4 units or more in Saint George's Respiratory Questionnaire (SGRQ) total score at Week 12

Ex-US APPROACH

- Time to first moderate or severe COPD exacerbation
- Time to first clinically important deterioration (CID) in COPD
- Change from baseline in average daily rescue Ventolin HFA use over 24 weeks
- Percentage of subjects achieving an MCID of 4 units or more in SGRQ total score over 24 weeks
- Change from baseline in the Exacerbations of Chronic Pulmonary Disease Tool (EXACT) total score over the treatment period
- Transient Dyspnea Index (TDI) focal score over 24 weeks

In both regulatory approaches, morning pre-dose trough FEV_1 for the attributable estimand will be considered a secondary analysis. Time to first moderate or severe COPD exacerbation in the subgroup of patients with a COPD exacerbation history of at least two moderate or severe exacerbations in the previous 12 months will also be considered a secondary analysis. Both analyses will be included in the Type I error control strategy as described in Section 6.4.8.

2.2.3 Other Efficacy Endpoints

Unless already categorized as a secondary endpoint in one of the regulatory approaches above, all of the following endpoints will be categorized as "Other efficacy endpoints".

- Rate of moderate or severe COPD exacerbations over the treatment period

- Change from baseline in morning pre-dose trough FEV_1, forced vital capacity (FVC), peak expiratory flow rate (PEFR), and forced expiratory flow between 25% and 75% of FVC (FEF_{25-75}) over 12, 24 and 52 weeks, and at each post-randomization in-clinic visit
- Rate of COPD exacerbations of any severity
- Time to first COPD exacerbation of any severity
- Rate of severe COPD exacerbations
- Time to first severe COPD exacerbation
- Rate of COPD exacerbations treated with systemic steroids
- Rate of COPD exacerbations treated with antibiotics
- Time to first COPD exacerbation treated with systemic steroids
- Time to first COPD exacerbation treated with antibiotics
- Time to first CID in COPD
- Time to first sustained CID in COPD
- Time to treatment failure (treatment discontinuation for any cause, moderate or severe exacerbation, or death)
- Time to death, all cause
- Time to death, respiratory
- Change from baseline in Exacerbations of Chronic Pulmonary Disease Tool (EXACT) total score, Evaluating Respiratory Symptoms in COPD (E-RS: COPD) total score (RS-Total Score), and symptom domain scores for breathlessness, cough and sputum, and chest symptoms over 24 weeks and 52 weeks, and over each 4-week interval of the 52-week treatment period
- Change from baseline in average daily rescue Ventolin HFA use over 12, 24, and 52 weeks and over each 4-week interval of the 52-week period
- Percentage of days with "no rescue Ventolin HFA use"
- TDI focal score over 12, 24 and 52 weeks, and at each post-randomization in-clinic visit
- Individual components of the TDI (functional impairment, magnitude of task, and magnitude of effort) over 12, 24 and 52 weeks, and at each post-randomization visit
- Percentage of subjects achieving an MCID threshold of 1 unit or more on average in TDI focal score over 12, 24 and 52 weeks
- Change from baseline in SGRQ total score over 12, 24 and 52 weeks, and at each post-randomization in-clinic visit
- Change in individual domains of the SGRQ (Symptoms, Activity, and Impacts) over 12, 24 and 52 weeks, and at each post-randomization in-clinic visit
- Percentage of subjects achieving an MCID of 4 units or more on average in SGRQ total score over 12, 24, and 52 weeks, and at each post-randomization in-clinic visit

- EuroQoL (EQ-5D) Dimensions Questionnaire (EQ-5D-5L) variables including the EQ-5D index score, the EQ-5D Visual Analog Score (VAS), and five dimension single item 5-level responses at each post-randomization in-clinic visit

2.2.4 Safety Endpoints

The safety endpoints for this study include:

- Adverse events (AEs), Treatment-emergent AEs, serious AEs (SAEs), AEs of Special Interest (AESIs)
- 12-lead electrocardiograms (ECGs)
- Clinical laboratory values (hematology and clinical chemistry)
- Vital signs measurements (blood pressure and heart rate)

2.2.5 Health Care Resource Utilization Endpoints

- The number of days missed from work due to COPD
- The number of days that primary caregivers of subjects missed from work as a result of the subject's COPD
- The percentage of subjects with telephone calls to health-care providers
  - Calls to any health-care provider (physician or other)
  - Calls to physician
  - Calls to other health-care provider
- The mean number of telephone calls to health-care providers
  - Calls to any health-care provider (physician or other)
  - Calls to physician
  - Calls to other health-care provider
- The percentage of subjects with visits to health-care providers
  - Visits to any health-care provider (general practitioner [GP], specialist, or other)
  - Visits to GP
  - Visits to specialist
  - Visits to other health-care provider
- The mean number of visits to health-care providers
  - Visits to any health-care provider (GP, specialist, or other)
  - Visits to GP
  - Visits to specialist
  - Visits to other health-care provider

- The percentage of subjects with Emergency Room (ER) visits
- The mean number of visits to ERs
- The percentage of subjects hospitalized
- The mean number of subject hospitalizations
- The mean number of days in the hospital
- The mean number of hospitalizations in which subjects spent some time in the Intensive Care Unit (ICU) or the Coronary Care Unit (CCU)
- The percentage of subjects hospitalized with some time spent in the ICU or CCU
- The mean number of days in the hospital with some time spent in the ICU or CCU
- The mean number of hospitalizations in which subjects spent no time in the ICU or CCU
- The percentage of subjects hospitalized with no time spent in the ICU or CCU
- The mean number of days in the hospital with no time spent in the ICU or CCU
- The mean number of days in the ICU
- The percentage of subjects in the ICU
- The mean number of days in the CCU
- The percentage of subjects in the CCU
- The percentage of subjects who required ambulance transport
- The mean number of times ambulance transport was required

3. STUDY DESIGN AND ANALYTICAL CONSIDERATIONS

3.1 Study Design

3.1.1 Overall Study Design and Plan

This is a Phase III, multi-center, randomized, double-blind, parallel-group, variable length efficacy and safety study comparing BFF MDI (320/9.6 µg and 160/9.6 µg) to FF MDI 9.6 µg administered BID, in subjects with moderate to very severe COPD. Eligible subjects must have at least 1 moderate or severe COPD exacerbation in the previous 12 months, and remain symptomatic, as measured by the COPD Assessment Test (CAT), while receiving one of more inhaled maintenance bronchodilators.

Subjects meeting inclusion criteria and no exclusion criteria will be randomized into this study. Randomization will be stratified by exacerbation history (1 or ≥ 2 moderate or severe exacerbations), post-bronchodilator FEV_1 (< 25% to < 50% predicted or 50% to < 80% predicted, measured at Visit 2), blood eosinophil (EOS) count (< 150 or ≥ 150 cells per mm^3), and country. Enrollment will be targeted to achieve a 2:1 ratio for the blood eosinophil strata with twice as many randomized subjects in the ≥ 150 cells per mm^3 category.

Following randomization, subjects will enter the Randomized Treatment Period. The study is variable in length, with a planned minimum of 12 weeks and a maximum of 52 weeks on

randomized treatment. The study will end when the last remaining randomized subject completes 12 weeks on randomized treatment or completes the Final Study Visit. The End of Study is defined as the date on which data are collected for the last subject's Follow-up Telephone Call.

This study will be conducted at approximately 275 sites, contributing approximately 6 to 10 subjects per site. A total of 1860 subjects will be randomized in a 1:1:1 scheme to one of the three treatment groups (see Table 1):

- BFF MDI 320/9.6 µg BID (620 subjects)
- BFF MDI 160/9.6 µg BID (620 subjects)
- FF MDI 9.6 µg BID (620 subjects)

Approximately 1581 subjects are expected to complete the study on randomized treatment to Week 12.

Table 1 Planned Sample Size

	BFF MDI 320/9.6 µg	BFF MDI 160/9.6 µg	FF MDI 9.6 µg	Overall
Allocation Ratio	1	1	1	
Randomized Subjects	620	620	620	1860
Subjects Completing the Study on Randomized Treatment to Week 12	527	527	527	1581

Subjects who discontinue randomized treatment prior to end of study will be encouraged to remain in the study to complete all remaining study visits. Subjects who agree to continue to be followed post treatment discontinuation will sign an informed consent form (ICF) addendum. All subjects who agree to continue study participation beyond treatment discontinuation will complete a Treatment Discontinuation/Study Withdrawal Visit prior to transitioning back to regularly scheduled study visits. Treatment discontinuation subjects will return to appropriate maintenance COPD medications, per the Investigator's discretion. For subjects recorded as Treatment Discontinuations that do not complete at least one post-treatment data collection, a telephone follow-up call is required at least 14 days after last study drug dose. All AEs/SAEs will be collected through the14 day follow up telephone call.

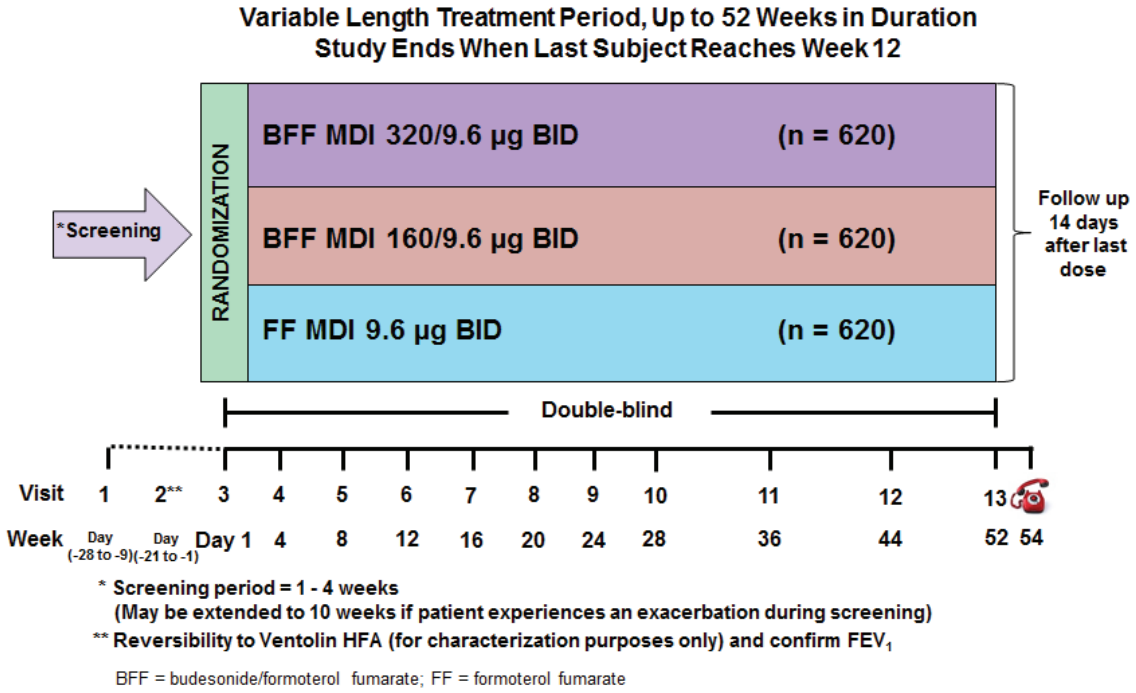
If a subject chooses not to continue with study assessments, at a minimum the subject will complete the Treatment Discontinuation/Withdrawal Visit (refer to the Schedule of Events in the Study Protocol). These subjects will return to appropriate maintenance COPD medications, per the Investigator's discretion. A follow-up telephone call will be performed at least 14 days after the last study drug dose. All AEs/SAEs will be collected through the 14 day follow up telephone call. In the event the Treatment Discontinuation/Withdrawal Visit is performed >14 days post last study drug dosing, a follow-up telephone call (TC) will not be required. These subjects will

be followed for vital status within 4 to 6 weeks after the last subject has been randomized into the study in accordance with the informed consent.

The Schedules of Events and Timed Assessments are in the study protocol.

The overall study design is summarized and illustrated in Figure 1.

Figure 1 Study Design



3.1.2 Prior, Concomitant, Post-Treatment, Prohibited Medications, and Other Restrictions

All prescription and over-the-counter (OTC) medications, as well as any herbal or vitamin supplements taken by the subject within 30 days before Visit 1 (Screening) will be recorded on the Prior/Concomitant Medications electronic Case Report Form (eCRF) with indication, total daily dose, dose regimen, and dates of drug administration. Refer to the Protocol for information about prohibited medications and other restrictions.

3.2 Hypothesis Testing

For the primary comparisons, the general null hypothesis will be that the mean treatment difference between BFF MDI and FF MDI is zero (mean treatment effects are equal). The alternative two-sided hypothesis is that the mean treatment difference is greater or less than zero (mean treatment effects are not equal).

The primary null (H_0) and alternative (H_1) hypotheses, with μ representing the mean, are presented below. The superiority hypotheses will take the following forms:

- H_0: $\mu_{BFF\ 320/9.6} = \mu_{FF\ 9.6}$
  H_1: $\mu_{BFF\ 320/9.6} \neq \mu_{FF\ 9.6}$

- H_0: $\mu_{BFF\ 160/9.6} = \mu_{FF\ 9.6}$
  H_1: $\mu_{BFF\ 160/9.6} \neq \mu_{FF\ 9.6}$

Two-sided P-value and 95% confidence interval (CI) for $\mu_{BFF} - \mu_{FF}$ will be computed.

3.3 Interim Analysis

No interim efficacy analyses are planned for this study.

The Data Monitoring Committee (DMC) will review safety data approximately every 6 months. Further detail is given in the DMC Charter.

3.4 Sample Size

It is estimated that a sample size of 1860 subjects (620 per arm in the BFF MDI and FF MDI groups) will provide power estimates for morning pre-dose trough FEV_1 as summarized in Table 2, and the power of detecting a numerical trend in the analysis of time to first moderate or severe COPD exacerbation as outlined in Table 3.

For morning pre-dose trough FEV_1, assumptions regarding variability are based on Pearl's experience with Phase IIb and III clinical studies. The expected standard deviation (SD) for the change from baseline at each visit is 200 mL. The expected SD over 24 weeks is 158 mL with a correlation of 0.55 over the six post-randomization visits to Week 24.

For the analysis of morning pre-dose trough FEV_1 at Week 12 (the US Approach), the proposed sample size of 1,860 subjects with 15% drop-out will provide approximately 90% power to detect a difference of 40 mL between BFF MDI and FF MDI. The Type I error will be controlled at a two-sided alpha level of 0.05.

For the analysis of morning pre-dose trough FEV_1 over 24 weeks (the ex-US Approach) with 30% dropout, the same proposed sample size will provide approximately 96% power to detect a difference of 40 mL between BFF MDI and FF MDI. The Type I error will be controlled at a two-sided alpha level of 0.05.

Table 2 Power Estimates for Morning Pre-Dose Trough FEV_1

Endpoint	Parameter	At Week 12 (US)	Over 24 Weeks (Ex-US)

Change from Baseline in Morning Pre-dose Trough FEV_1	Assumed Difference (BFF – FF)	40 mL	40 mL
	Assumed Standard Deviation for the Difference	200 mL	158 mL
	Assumed Correlation over the Six Post-randomization Visits to Week 24	-	0.55
	Dropout Rate	15%	30%
	Power	90%	96%

Assumptions regarding the numerical trend for the observed hazard ratio (HR) between BFF MDI and FF MDI are based on a literature review and data from a recent COPD exacerbation study of a budesonide and formoterol fumarate combination therapy. The percentage of subjects with at least one moderate or severe COPD exacerbation in the BFF MDI 320/9.6 µg and FF MDI 9.6 µg groups is estimated to be 28.3% and 34.0%, respectively. This represents a hazard ratio of approximately 80% for BFF MDI compared to FF MDI. The projected numbers of subjects with at least one moderate or severe COPD exacerbation in this variable length study are 195 and 234, respectively, for the BFF MDI 320/9.6 µg and FF MDI 9.6 µg groups. This would provide 64% power to detect an HR of 80% or lower. The probability of detecting a numerical trend in the analysis of time to first moderate or severe COPD exacerbation is estimated to be 87%, where numerical trend is defined as the observed HR between a BFF MDI dose and FF MDI being below 0.90. If the criterion for numerical trend is raised to 0.95, the probability of detecting a numerical trend is estimated to be 94%. All calculations assume Type I error control at a 2-sided alpha level of 0.05.

Table 3 Power Estimates for Incidence of Moderate or Severe COPD Exacerbations

Endpoint	Parameter	Over the Treatment Period
Time to first Moderate-Severe COPD Exacerbation	Assumed BFF MDI COPD Exacerbations incidence	28.3% of subjects
	Assumed FF MDI COPD Exacerbations incidence	34.0% of subjects
	Assumed Hazard Ratio (BFF/FF)	0.8
	Power	64%

	Probability to detect numerical trend: estimated HR ≤ 0.9	87%
	Probability to detect numerical trend: estimated HR ≤ 0.95	94%

4. DATA AND ANALYTICAL QUALITY ASSURANCE

The overall quality assurance procedures for the study data, statistical programming and analyses are described in Standard Operating Procedures (SOPs) of [redacted]. Detailed data management procedures are documented in the study Data Management Plan, Data Validation Check Specifications, and Integrated Safety Data Review Plan. Detailed statistical and programming quality control and quality assurance procedures are documented in the Statistical Analysis and Programming QC/QA Plan.

Transfer of Pulmonary Function Test (PFT) data from the central PFT laboratory [redacted] will be defined in the [redacted] DMP (Data Management Plan), and data handling rules related to this data are included in Appendix 1 of this SAP. The quality of all PFT's obtained at each time point will be graded independently at [redacted] by qualified personnel. Quality grading assessments will be based on American Thoracic Society/European Respiratory Society (ATS/ERS) criteria and will be included in data transfers.

5. ANALYSIS POPULATIONS

5.1 Population Definitions

5.1.1 Intent-to-Treat (ITT) Population

The **Intent-To-Treat Population** is defined as all subjects who are randomized to treatment and receive any amount of the study treatment. Subjects will be analyzed according to the treatment they were assigned at randomization. Data obtained after discontinuation of treatment, but prior to withdrawal from the study, will be included. The ITT population will be used for sensitivity analyses.

5.1.2 Modified Intent-to-Treat (mITT) Population

The **Modified Intent-to-Treat Population** is a subset of the ITT Population, defined as all subjects with post-randomization data obtained prior to discontinuation from treatment. Any data collected after completion of or discontinuation from randomized study medication will be excluded. Subjects will be analyzed according to randomized treatment group. (Note that a subject who used a study treatment, but took less than one full dose of treatment will qualify for this population). The mITT Population will be the primary population for all efficacy analyses. Note: The knowledge that a subject did not have a COPD exacerbation constitutes an efficacy assessment.

5.1.3 Rescue Ventolin User Population

Differences in rescue Ventolin HFA usage are expected across the study including some subjects who used virtually no rescue medication at study entry. In order to represent the population of patients who may benefit from study treatment and reduce their use of rescue medication, the Rescue Ventolin User (RVU) Population is defined as all subjects in the ITT Population with average baseline Rescue Ventolin use of ≥ 1.0 puff/day.

5.1.4 Safety Population

The **Safety Population** is defined as all subjects who are randomized to treatment and receive at least one dose of the study treatment. Subjects will be analyzed according to treatment received rather than randomized. If a subject received more than one randomized treatment, they will be analyzed and included in summaries according to the treatment they received the most. Subjects receiving no study treatment will be excluded, as will subjects who have no post-dose safety assessments. A subject who used a study treatment, but took less than one full dose of treatment will qualify for this population. Note: The statement that a subject had no AEs also constitutes a safety assessment.

5.2 Populations for Primary and Sensitivity Analyses

Demographics will be summarized for the mITT, RVU, Safety, and Non-Randomized Populations.

Extent of exposure will be summarized for the Safety Population. The Safety Population will be used to summarize safety..

Efficacy analyses will be performed for the ITT and mITT Populations. In general, the mITT Population will be considered the primary population for the efficacy analyses, with the ITT population being considered supportive. Rescue medication endpoints will be analyzed with mITT, ITT, and RVU populations.

A Per Protocol population is not defined in this study. However, important protocol deviations will be identified according to the Statistical Protocol Deviation Plan (sPDP) and blinded data review meeting (BDRM) process. These deviations will be summarized (*Table 1.3.1*) and listed (*Listing 1.6.2*).

6. STATISTICAL ANALYSIS

Analyses will be performed when the final database is available. All data collected contributing to the analysis will be provided in listings. Data for all subjects who are randomized will be included in the subject data listings. Data for non-randomized subjects will be listed where available.

All safety and efficacy parameters will be summarized by treatment unless specified otherwise.

Continuous variables will be summarized with descriptive statistics (the number of non-missing values, mean, standard deviation, median, minimum, and maximum). Additionally, the 25th and 75th percentiles will be presented when appropriate based on historical knowledge of the normality or non-normality distribution of underlying data.

Categorical variables will be summarized with frequency counts and percentages (where appropriate).

6.1 Data Handling Rules and Definitions, Including Handling of Missing Data

Missing data will be maintained as missing in the analysis datasets, unless specified otherwise. For variables where missing data are imputed, the analysis dataset will contain a new variable with the imputed value and the original variable value will be maintained as missing.

Data Imputation for Adverse Events Summaries by Severity and Relationship to Study Drug

For the AE summaries by severity (mild, moderate, or severe), an AE with missing severity will be deemed as severe. For the AE summaries by relationship to study drug, an AE with a missing relationship to study drug will be deemed as related. Imputed values will not be listed in data listings.

Use of Data from Unscheduled Assessments for Laboratory, Vital Sign, and ECG Summaries (Continuous Parameters)

Data from unscheduled visits will not be used for by-visit summaries. Data from both scheduled and unscheduled visits will be used for shift tables and for determining incidence of clinically significant values.

Data Imputation (All Laboratory Summaries)

Laboratory values of '>=x' or '<=x' will be taken as the value of x in the analyses. If a laboratory value is prefixed with'>': the available original value +0.001 will be used for table summaries; if a laboratory value is prefixed with '<', then the original value –0.001 will be used in table summaries.

Study Dates and Day of Assessment or Event

Study Day and Day of Assessment or Event definitions are provided in Appendix 1, Data Handling Rules.

Pre-dose spirometry values will use the average of the non-missing -60 minutes and -30 minutes values. Weekly averages for eDiary-based (where eDiary means electronic diary) parameters will use all non-missing values.

On-treatment COPD exacerbations

An exacerbation will be considered "on-treatment" if its start date is before or on the last treatment date. For treatment discontinuations, this definition is extended to include exacerbations starting one day after the last treatment date. (If it is decided during a clinic visit to discontinue study drug and to switch to a treatment for the ongoing exacerbation symptoms, the subject typically would not take the morning dose of study drug at that visit, and their exacerbation start date will be one day after the last treatment date. Such exacerbations will still be considered "on-treatment").

6.2 Subject Disposition and Analysis Populations

Disposition for all randomized subjects will be tabulated *(Table 1.1.1)* and listed *(Listing 1.2).* The tabulation will include the number of subjects in each randomized treatment who were not treated, who received the study treatment, who discontinued treatment prematurely, who withdrew from the study prematurely, and who completed the study. The number and percentage of randomized subjects included in the mITT, ITT, Rescue Ventolin User, and Safety Populations will also be tabulated *(Table 1.1.1)*. Informed consent is listed in *Listing 9.6*.

The numbers of subjects randomized and in the analysis populations will be provided by country, center, and treatment in *Table 1.1.2*. The number of subjects randomized by stratification factor based upon clinical data will be tabulated in *Table 1.1.4*. The duration of treatment and study participation per subject will be summarized in *Table 1.1.5*. If there are any subjects who took study treatment other than what was randomized during the study, both the treatment assigned at randomization and actual treatment(s) received during the Treatment Period will be listed *(Listing 1.3)*. The duration of actual treatment and duration of study participation will also be listed (*Listing 1.3)*. A list of subjects with discrepant IWRS-based and actual stratification factors will also be provided (*Listing 1.7*).

A summary of reasons subjects were not randomized will be provided for all subjects not randomized (*Table 1.1.3*). A listing of reasons subjects were not randomized will also be provided *(Listing 1.4)*. Subjects excluded from the ITT, mITT, RVU, and Safety analysis populations will be listed (*Listing 1.6.1)* for all subjects randomized. Reasons for premature discontinuation from study treatment will be summarized for the Safety Population *(Table 1.2.1).* Similarly, reasons for subjects' withdrawal from the study will be summarized for the ITT Population (*Table 1.2.2*).

Time to discontinuation of treatment and withdrawal from the study will be presented graphically by means of the Kaplan-Meier plots (*Tables and Figures 1.2.3 and 1.2.4*).

Important protocol deviations will be listed and summarized for all subjects randomized *(Table 1.3.1 and Listing 1.6.2)*. A listing of subjects who did not comply with restrictions on smoking, use of rescue medication, and xanthine-containing products prior to spirometry (Section 5.1.4.1) will be provided in *Listing 6.1.1*. Use of rescue medication at pre-dose or during the post-dose assessments on each specific test day (yes/no) will be tabulated in *Listing 6.1.3*. In addition, the eligibility information (inclusion/exclusion criteria with any waivers granted) of all subjects who are randomized will be listed *(Listing 2.1).*

The number and percentage of subjects with changes in smoking status after the start of study treatment will be tabulated by randomized treatment, by visit and overall during the study in *Table 1.13 (Safety Population)* and listed *(Listing 1.5)*.

6.3 Demographic and Baseline Characteristics and Extent of Exposure

The definitions for the derived demographic or baseline characteristic variables can be found in Appendix 1.

6.3.1 Demography, Baseline Characteristics, and CAT

Subject demographics, total CAT score, use of inhaled corticosteroids (ICS) at screening, and smoking status/history will be summarized for the mITT, RVU, and Safety Populations and for Non-Randomized subjects *(Tables 1.4.1 to 1.4.4,* respectively*, and Listing 1.2)*. The ITT population does not need to be tabulated because it is the same as the mITT population for demographics and baseline characteristics. If the Safety Population has the same treatment assignment as the mITT, then these summaries will be identical as well and hence not produced. Inhaled corticosteroid use (yes/no) will be summarized for all populations except for the Non-Randomized subjects.

Demographic and baseline characteristic variables summarized will include the following:

- Age
- Age Group
- Age of onset of COPD
- Gender
- Race
- Ethnicity (Hispanic or Non-Hispanic)
- COPD Assessment Test (CAT) total score and total score category (<10, ≥10, <15, ≥15, <20, ≥20, Missing)
- Used inhaled corticosteroids at Screening (all populations except for Non-Randomized subjects)
- Baseline eosinophil count (<150 cells per mm^3 vs. ≥150 cells per mm^3)
- Baseline exacerbation history (group as 1 and ≥2 M/S COPD exacerbations in the previous 12 months)
- Smoking status (current vs. former smoker)
- Number of years smoked
- Average number of cigarettes smoked per day
- Number of pack years smoked, calculated as (number of cigarettes per day/20) x number of years smoked
- Weight
- Height
- Body mass index (BMI)

Screening and pre-treatment CAT data will be listed *(Listing 4.2)*.

6.3.2 COPD History, Screening/Baseline Spirometry, and Reversibility

Duration of COPD: the number of years prior to the start of study medication that COPD was first diagnosed (calculated as [Date of First Dose of Study treatment in the study – Date COPD First Diagnosed] /365.25) will be summarized by treatment and all subjects for the mITT and Safety Populations and listed *(Tables 1.5.1, 1.5.2 and Listing 4.1)*. The summary for the Safety Population will only be presented if it is different from the mITT/ITT Population summary. Severity of COPD at Screening Visit 2 post-Ventolin HFA will also be included in these summaries. History of moderate or severe COPD exacerbations within the past 12 months will be summarized and listed for subjects in the mITT and Safety Populations *(Table 1.9.1, 1.9.2, and Listing 4.3)*.

Descriptive statistics will be provided for screening period pre-bronchodilator and post-bronchodilator and baseline spirometry parameters *(Tables 1.6.1 to 1.6.2* for the mITT and RVU Populations, respectively, and *Listing 2.2)*.

Characterization of Reversibility:

Reversibility to Ventolin HFA will be evaluated at Visit 2 and used as a stratification variable at randomization to ensure an even distribution of reversibility across the treatment arms. A subject is considered reversible if the improvement in FEV_1 at 30 minutes post-Ventolin is ≥ 12% and ≥ 200 mL.

Reversibility to Ventolin HFA at Screening Visit 2 will be summarized for the mITT Population and listed *(Table 1.7.1, Listing 2.2 and Listing 5.2* for Ventolin HFA dispensing). The number and percentage of subjects reversible will be included in these summaries. Also included will be a summary of the change in FEV_1 from pre-dose FEV_1 to post-bronchodilator assessment. If multiple time points are available post-bronchodilator, then the one with the highest FEV_1 will be used.

Additionally, the number and percentage of subjects meeting each of the following response criteria will be summarized for Ventolin HFA bronchodilator:

≥12% improvement post-bronchodilator in FEV_1 from pre-bronchodilator

≥150 mL improvement post-bronchodilator in FEV_1 from pre-bronchodilator

≥200 mL improvement post-bronchodilator in FEV_1 from pre-bronchodilator

6.3.3 Medical and Surgical History at Screening, Reproductive Status and Pregnancy Testing

Medical and Surgical History at Screening will be summarized for the Safety Population and listed for all randomized subjects *(Table 1.8.1.1 and Listing 4.4)*. Cardiovascular medical history of interest at Screening will be summarized for the Safety Population and listed for all randomized subjects *(Table 1.8.1.2 and Listing 4.5)*.

Screening Reproductive Status and Pregnancy Testing Results will be listed *(Listing 4.6)*.

6.3.4 Prior, Concomitant, and Post-Treatment Medications/Treatments

All prescription and OTC medications, as well as any herbal or vitamin supplements, taken by the subject within 30 days of Visit 1 and all concomitant therapy taken by the subject while on study will be recorded on the Prior and Concomitant Medications case report form (CRF) page.

Coding: Verbatim medication/treatment terms will be coded by [redacted] and will be assigned a preferred term and an ATC (anatomic therapeutic class) term using the latest version of the World Health Organization Drug Dictionary (WHO-DD) available (version: 3Q2016 or later).

Multiple ATC assignments: If there are multiple ATC codes assigned to the same concomitant medication, the "primary" one based on a Pearl medical evaluation will be used.

Prior medication/treatment is any medication/treatment taken prior to study treatment, even if this medication continued to be taken on the day of the start of study treatment in the study or afterward *(Appendix 1)*.

Concomitant medication/treatment is any medication/treatment reported as being taken after the start of the randomized study treatment in the study to the date prior to the last dose of study treatment for the subject. A medication with an onset date on or after the date of discontinuation from or completion of randomized study treatment for the subject will not be considered concomitant, but will be considered a **Post-Treatment medication/treatment**.

Any medication/treatment which cannot be identified as Prior, Concomitant, or Post-Treatment will be considered as being in each of the categories that are possible from the available information.

Concomitant/post-treatment COPD, COPD-Exacerbation, and Non-COPD related medications/treatments will be summarized by preferred term and actual treatment received for the Safety Population *(Tables 1.11.1 to 1.11.7)*. COPD-related summaries will not include the COPD-exacerbation medications. Prior, concomitant/post-treatment COPD, COPD-Exacerbation, and Non-COPD medications will be displayed in separate listings *(Listings 4.7 to 4.9,* respectively*)*.

Reported prior medications for COPD, COPD-Exacerbation, and non-COPD-related medications will be tabulated for the Safety Population *(Tables 1.10.1 to 1.10.4)* and listed separately *(Listings 4.7 to 4.9, respectively)*.

Prior COPD medications will be tabulated (for the Safety population) for subjects having received any one, two, all three, or none of the following treatments (whether in fixed combination products or separately): (1) a long-acting muscarinic antagonist (LAMA), (2) a long-acting β2 agonist, and (3) an inhaled corticosteroid *(Table 1.10.2)*. For this purpose, scheduled SAMA (Short-acting muscarinic antagonist) or SABA treatments are included. In addition, tabulations for long-acting muscarinic antagonists (LAMA) and long-acting β2 agonists (LABA) will also be included.

Post-treatment medications will be tabulated for subjects having received any one, two, all three, or none of the following treatments: (1) a muscarinic antagonist, (2) a β2 agonist, and (3) an ICS *(Table 1.11.5)*.

6.3.5 Extent of Exposure to Study Medication and Compliance

Subject's exposure to a study treatment will be determined by the duration of time (days) for which the doses were administered, defined as "([End date of treatment – Date of first dose of treatment] + 1)". Percent compliance is defined as (total number of puffs of study treatment taken on a study day/total expected puffs taken on a study day) averaged across all days of a subject's dosing between start of study treatment and last day on study treatment) x 100.

The expected number of puffs for a test day which is the last date of treatment will be 2, and the expected number of puffs for the last date of treatment which is not a test day will be 4 when a PM dose is taken but will be 2 otherwise; the expected number of puffs on dates prior to the last date of treatment will be 4.

The number of days of exposure to study treatment will be summarized for each treatment for the Safety Population. The total person-years of exposure for a treatment group, defined as the total exposure in the study across all subjects in the treatment, will also be provided by treatment *(Table 1.12)*. In addition, treatment compliance will be provided in this summary. The treatment compliance will be categorized into 7 different groups depending on the degree of compliance: 0 – <20%, ≥20 – <40%, ≥40 – <60%, ≥60 – <80%, ≥80 – ≤100%, >100 – ≤120%, and >120%. Also provided in this summary will be descriptive statistics (n, mean, standard deviation, median, minimum and maximum) for percent compliance by treatment. Treatment compliance will be reported in *Listing 5.3*. A listing of treatment dosing and dispensing information will be provided in *Listing 5.1.* Any comments related to study medication or any other additional study comments will be listed *(Listing 9.6)*.

6.4 Efficacy Analyses

6.4.1 Estimands

Three estimands are defined for this study: the efficacy estimand, attributable estimand, and treatment policy estimand.

The primary estimand of interest is the efficacy estimand and is the effect of the randomized treatments in all subjects assuming continuation of randomized treatments for the duration of the study regardless of actual compliance. There are two additional estimands of interest. One is the attributable estimand, which is the effect of treatment in subjects attributable to the randomized treatment. For this estimand, discontinuation of randomized medication for reasons such as tolerability or lack of efficacy is considered a bad outcome. The second additional estimand of interest is the treatment policy estimand. This estimand is the effect of randomized treatment over the study period regardless of whether randomized treatment is continued.

The primary analysis for the efficacy estimand will be conducted using the mITT Population where only data obtained prior to subjects discontinuing from randomized treatment will be

utilized. This assumes that efficacy observed on treatment is reflective of what would have occurred after discontinuation of randomized treatment had they remained on treatment.

The second estimand of interest is the attributable estimand. Analyses of the attributable estimand will be conducted in the mITT Population. Data that are missing due to treatment discontinuation will be imputed based on the 5^{th} or 95^{th} percentile of the reference arm's (FF MDI) distribution if the reason is reasonably attributable to tolerability or lack of efficacy. The 5^{th} percentile applies to an endpoint for which a higher value is a better outcome; however the 95^{th} percentile applies to an endpoint for which a higher value is a worse outcome. Other missing data are to be imputed using the observed data model, i.e. assumed to be missing at random (MAR). The number of imputations used for the derivation of the attributable estimand will be between 100 and 1000. More detail about the computation of the attributable estimand will be provided in subsequent sections (especially 6.4.4.1) and in the Details Appendix of this SAP.

Treatment discontinuations reasonably attributable to tolerability or lack of efficacy will be identified during the BDRM and documented in the BDRM minutes prior to unblinding. Discontinuations will be attributed to tolerability if the subject had an adverse event determined by the Investigator to be related to study drug, and for which study drug was permanently discontinued. Discontinuations will be attributed to lack of efficacy if 'lack of efficacy' is indicated to be the primary reason for discontinuation from study drug. For the remaining discontinuation categories, where specific reasons or criteria frequently need to be considered, decisions will be made and documented at the BDRM.

The third estimand of interest is the treatment policy estimand. Analyses of the treatment policy estimand will be conducted in the ITT Population, in which all observed data will be utilized regardless of whether subjects remain on randomized treatment.

There are two pairwise comparisons of interest, namely, BFF MDI (2 doses: 320/9.6 µg and 160/9.6 µg) vs. FF MDI. Estimation results will be provided by randomized treatment and for each treatment difference for all comparisons, in each estimand. All comparisons will be performed for testing superiority.

6.4.2 Baselines and Baseline Covariates for Analysis

The mean of all evaluable 60- and 30-minute pre-dose spirometry assessments conducted at Day 1 (Visit 3) will be used to establish baseline for all FEV_1, FVC, $FEF_{25-75,}$ and PEFR parameters.

For the diary symptom score parameters and rescue medication usage, baseline will be the average of the non-missing values from the diary data collected in the last seven days of the Screening Period.

For the SGRQ scores, baseline will be the value of the score calculated using the Day 1 questionnaire data collected prior to the start of randomized study treatment.

Baseline COPD exacerbation history (from the Visit 1 CRF page) is set to 1 or ≥ 2 moderate-or-severe COPD exacerbations in the last 12 months.

ICS use at screening (Yes or No) is to be defined as follows - a subject will be considered to have had "ICS Use at Screening" if:

- the subject was taking a medication that contained a glucocorticoid component (active ingredient) that is listed in the WHODRUG SDG (standardized drug grouping) of "CORTICOSTEROIDS", and
- the route of administration was "INHALED", and
- the medication was used at any time during the screening period (or in the 30 days prior to the screening period).

Baseline blood eosinophil count is the average of non-missing blood eosinophil count values prior to the first dose of study medication.

Baseline age is the age in years at the time of Informed Consent.

Baseline post-bronchodilator FEV_1 is the highest available value of FEV_1 obtained after dosing with Ventolin at Visit 2.

Baseline percent reversibility to Ventolin is 100 x (POST-PRE)/PRE, where PRE is the mean of the available 30 minute and 60 minute values of FEV_1 prior to dosing with Ventolin at Visit 2, and POST is the post-bronchodilator FEV_1 value defined above.

Visits and Time Windows for Visit-Based Efficacy Assessments:

Efficacy data obtained during unscheduled visits will not be used for any of the pre-defined efficacy analyses. Efficacy from scheduled and unscheduled visits will be listed.

6.4.3 Impact of Variable Length of the Study on the Analyses

The study will end when the last remaining randomized subject completes 12 weeks on randomized treatment. Given the variable length nature of the study, the distinction between "actual last visit" and "planned last visit" will be important for the analyses where the imputation of missing values is required.

Unlike a fixed duration study where each subject can reasonably be expected to stay in the study for the protocol-specified treatment period, subjects in a variable-length study will likely have different "planned last visits", with those visits conditional on a protocol-specified event (in this case, the last subject being randomized). Section 8 of the protocol presents the rules for determining the "planned last visit" for subjects who are still in the study at the time of last subject randomized. For subjects who have withdrawn from the study before the last subject is randomized, their "planned last visit" will be determined based on the same rules, applied to the projected schedule of visits (*Listing 1.3*).

When required for an analysis, imputation of missing data will only be performed up to and including the subject's "planned last visit". There will be no imputation of data beyond the subject's "planned last visit" since those data were not expected to be collected per protocol. For analysis purposes, those data are not considered missing.

6.4.4 Primary Efficacy Analysis

The primary efficacy analysis of change from baseline in morning pre-dose trough FEV_1 will be conducted for the efficacy estimand.

6.4.4.1 Change from Baseline in Morning Pre-Dose Trough FEV_1

Change from baseline in morning pre-dose trough FEV_1 is the primary endpoint in the US approach at Week 12 and the Ex-US approach over 24 weeks. It is also the 'other' endpoint over 52 weeks and at each post-randomization in-clinic visit.

Change from baseline in morning pre-dose trough FEV_1 at each visit is defined as the average of the 60 and 30 minute pre-dose values minus baseline. In subjects missing either of these pre-dose assessments, the value will be calculated from the single measurement. In subjects missing both pre-dose values, morning pre-dose trough FEV_1 at that visit will not be calculated. Spirometry data from unscheduled visits will not be used for this analysis. Assessments obtained during early termination visits will be used if their timing is consistent with the next scheduled collection of spirometry data.

The change from baseline in morning pre-dose trough FEV_1 will be analyzed using a repeated measures (RM) linear mixed model. The model will include treatment, visit, and treatment-by-visit interaction, and ICS use at Screening as categorical covariates and baseline FEV_1, baseline blood eosinophil count, and percent reversibility to Ventolin HFA as continuous covariates. An unstructured covariance (UN) matrix will be used to model correlation within a subject. If this model fails to converge, then a first-order autoregressive (AR(1)) structure will be used instead. In AR(1) model, subject will be included as a random effect.

Contrasts will be used to obtain estimates of the treatment differences at Week 12 and over 24 and 52 weeks, and at each post-randomization visit. Two-sided p-values and point estimates with two-sided 95% CIs will be produced for each treatment difference. All comparisons will be for superiority.

The primary analysis will be conducted for the efficacy estimand. The analysis of the primary endpoint for the attributable estimand is considered secondary. Supportive analyses will be also conducted for the treatment policy estimand (*Tables* and *Figures 2.1.1 to 2.1.3* and *Figures 2.1.1.1 to 2.1.1.3* for the efficacy, attributable, and treatment policy estimands, respectively).

For the attributable estimand (for the analysis at Week 12 and the analysis over 24 weeks), multiple imputation for missing values for morning pre-dose trough FEV_1 will use mean changes from baseline based on the 5^{th} percentile of the reference arm's (FF MDI) distribution when missingness is reasonably attributable to tolerability or lack of efficacy (see Section 6.4.1). Other

missing data are to be imputed using the observed data model. The variance used for the multiple imputation are described in the Details Appendix to this SAP. The number of imputations used for the derivation of the attributable estimand will be between 100 and 1000. Work by Seaman, White and Leacy (2014) and Cro (2017) show that Rubin's rules can be validly used in conjunction with so called control-based multiple imputation methods, of which the attributable analysis is one type. Given these results we believe the attributable estimand analysis to be conservative from a Type I error control perspective.

Exploration of the robustness of findings to missing data is discussed in Section 6.4.4.3.

6.4.4.2 Assumptions Checks and Removal of Outliers in Sensitivity Analyses

In general the distribution of spirometry measures is well-approximated by a normal distribution. Under some circumstances, atypical values can arise. Such values may disproportionately affect model-based estimates of the fixed effect and variance parameters. Prior to database lock and unblinding, the change from baseline values for efficacy endpoints will be examined as part of data quality management. This may include production of normal probability plots, kernel density estimates, and normal order outlier statistics. Based on this blinded evaluation, if atypical values are identified, nonparametric methods or data transformations (e.g. logarithmic or normal rank transformation) will be considered. If erroneous values are detected, every effort will be made to correct them prior to database lock. If these values cannot be corrected, they will be considered for removal from analysis. These analyses will be conducted if warranted to demonstrate the robustness of the primary and secondary results and reported in the Statistical Methods Appendix to the clinical study report (CSR).

The assumption of normality for the change from baseline in the morning trough FEV_1 data will be checked by visually inspecting the distribution of the residuals. Also, model fit and the assumption of homogeneity of variance will be verified by inspection of scatter plots of predicted vs. residuals, residuals vs. treatment, residuals vs. ICS use (yes/no), and by box plots of residuals for model variables with a potential effect on variance (treatment, visit, and ICS use). Plots for scaled (marginal) residuals will be prepared (option=VCIRY on the model statement and ODS graphics option allows the production of plots using these residuals). As a sensitivity analysis, if appropriate, the linear RM model analysis will be conducted by allowing for heterogeneity of variance between treatments, visits (if unstructured covariance model fail to converge), and/or ICS use categories (yes/no). Note that the unstructured covariance structure allows for heterogeneity among the visits.

Some further assumptions checks are mentioned in the Details Appendix to this SAP.

6.4.4.3 Sensitivity Analyses for Missing Data

Sensitivity analyses will be conducted for the change from baseline in morning pre-dose trough FEV_1 to evaluate the robustness of the primary analysis findings to missing data.

Robustness of results to missing data will be explored using tipping point analyses (Ratitch 2013) for the efficacy, attributable and treatment policy estimands. The following table

summarizes the multiple imputation-based sensitivity analyses under the PMM (pattern mixture model) framework that will be undertaken.

Table 4 Sensitivity Analyses for Morning Pre-Dose Trough FEV_1

Efficacy Estimand		Attributable Estimand	Treatment Policy Estimand
mITT Population		mITT Population	ITT Population
Tipping point analysis #1: All missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values that are considered MNAR are imputed with the change from baseline in the treatment arm decremented by up to 500 mL until the p-value ≥ 0.05.	Tipping point analysis #2: All missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values are imputed with the change from baseline in the treatment arm decremented by up to 500 mL until the p-value ≥ 0.05.	Tipping point analysis: MI based on the 5th percentile of the reference arm's distribution if treatment discontinuation is due to tolerability or lack of efficacy of study drug (as in the primary analysis of this estimand). Otherwise, all missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm), values are imputed with the change from baseline in the treatment arm decremented by up to 500 mL until the p-value ≥ 0.05.	Tipping point analysis: All missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values are imputed with the change from baseline in the treatment arm decremented by up to 500 mL until the p-value ≥ 0.05.

MNAR = Missing not at random. MNAR will be defined and documented in the BDRM minutes prior to unblinding. The tipping point will be shown to at least a precision of 10 mL. Imputed values may not be impossible values – i.e. changes from baseline that would imply a negative FEV_1 value. Thus the values will be imputed from a truncated distribution.

The primary analysis is for the efficacy estimand that includes data collected up until the time of discontinuation of treatment. The efficacy estimand quantifies the difference in outcomes for all patients as if they continued on their initially randomized treatment. The primary analysis uses a linear mixed model and assumes that all missing data are MAR or MCAR (missing completely at random).

Although the analysis for the attributable estimand starts with the same amount of missingness, less remains after imputation for missingness deemed attributable to the treatments is performed. These remaining missing data are imputed using the observed data model in the main analysis under the assumption of MAR. More detail about the computation of the attributable estimand is provided in subsequent sections and in the Details Appendix to this SAP.

Tipping-point analyses will be conducted to examine the impact of varying the treatment mean for missing data in subjects who discontinue BFF MDI. Multiple imputation (MI) techniques will be used to impute the missing data for these patients by varying the mean in the treatment arm. The change from baseline in the treatment arm will be decremented by up to 500 mL until the p-value for the comparison of treatment to comparator becomes ≥ 0.05. A total of 10 imputations will be used for each set of tipping point analyses. This imputation technique will be applied in sensitivity analyses as described below.

Tipping Point Analyses of the Primary Estimand:

- Tipping Point #1: this first set of analyses will impute diminished effects only for subjects on BFF MDI whose missing data are determined to be MNAR.
- Tipping Point #2: this analysis will impute diminished effects for all missing data in the BFF MDI arm.

Note that for both tipping point analyses, all other missing data will be imputed using the observed data model.

Tipping Point Analysis of the Attributable Estimand:

For the attributable estimand, by definition, missing data in all arms due to tolerability and lack of efficacy are already imputed based on the 5th percentile of the reference arm's distribution, therefore the remaining missing data imputed using the observed data model in the main analysis are likely MAR or MCAR. Hence, there is no need to conduct a tipping analysis like #1 planned for the efficacy estimand. A tipping point analysis like #2 will be conducted where the non-attributable missing data will be imputed using progressively diminished effects.

Tipping Point Analysis of the Treatment Policy Estimand:

For the treatment policy estimand, a tipping point analysis like #2 will be conducted where missing data in the treatment arm will be imputed using progressively diminished effects.

In all of these analyses, the imputed values that would have been seen are then combined with the observed values to provide a complete dataset. These data are then analyzed using the same linear mixed model used for the primary analysis. This analysis is repeated multiple times and the results are combined using Rubin's formulae [Rubin, 1987].

For the tipping point analyses, tables giving results for each progressively diminished effect will be produced. Figures of delta (decrement in treatment effect) versus p-values will also be produced. Details of the sensitivity analyses are discussed in the Statistical Methods Appendix to the CSR.

6.4.4.4 Cumulative Responder Analysis

Additional sensitivity analyses will be implemented based on a cumulative responder approach (Farrar et al., 2006) for the change from baseline in morning pre-dose trough FEV_1 at 12 Weeks

and over 24 weeks *(Tables 2.1.6.1 and 2.1.6.2* for the efficacy estimand, *Tables 2.1.6.3* and *2.1.6.4* for the treatment policy estimand). For the purpose of this analysis only, the change from baseline in morning pre-dose trough FEV_1 over 24 weeks will be defined at the patient level as the simple average of all available values of the change up to Week 24.

A cumulative distribution plot by treatment arm will also be produced. The observed change from baseline in morning pre-dose trough FEV_1 at Week 12 and over 24 weeks will be plotted on the x-axis, while the proportion of responders (subjects that equal or exceed that level of change) will be plotted on the y-axis *(Figures 2.1.6.1* and *2.1.6.2* for the efficacy estimand*, Figures 2.1.6.3* and *2.1.6.4* for the treatment policy estimand). Subjects without post-baseline data will be considered non-responders in the analysis. For display purposes only, the range of the X axis will be from -1 to +1 liters [L] by increments of 0.01 liters in order to avoid the undue influence of outlying values. The cumulative responder curves for each treatment will then be compared pairwise using Kolmogorov-Smirnov tests. A cumulative responder analysis for the attributable estimand will not be performed as methodology to apply Rubin's rules for combining multiply imputed data for such an analysis is not readily available.

6.4.5 Secondary Efficacy Analyses

Secondary efficacy analyses will be conducted for the efficacy, attributable, and treatment policy estimands. The efficacy estimand will be considered the primary. The analysis of the primary endpoint with the attributable estimand is considered a secondary efficacy analysis. Otherwise, the attributable estimand will be supportive. The treatment policy estimand will be supportive.

6.4.5.1 Time to First Moderate or Severe COPD Exacerbation

Time to first moderate or severe COPD exacerbation is a secondary endpoint in the US and the Ex-US approaches.

Time to first moderate or severe COPD exacerbation is the time from first dose of study medication to the time of onset of the first moderate or severe COPD exacerbation.

Only on-treatment exacerbations will be included for calculating the time to first moderate or severe COPD exacerbation for the efficacy estimand (see section 6.1).

Exacerbations occurring after the premature discontinuation of treatment will be considered for the treatment policy estimand.

For the attributable estimand, missing data in all arms due to tolerability and lack of efficacy will be imputed using the 95^{th} percentile of the reference arm's (FF MDI) distribution. The imputed value will be drawn from a negative binomial distribution with mean exacerbation rate (and variance) based on the 95^{th} percentile of the reference arm's distribution. The remaining missing data will be imputed using the observed data model, i.e. assumed to be missing at random (MAR) or missing completely at random (MCAR). Further information about the computation of the attributable estimand is described in the Details Appendix to this SAP (Appendix 6); however, the timing of imputed events is also needed. These will be obtained for each imputed

event by randomly drawing a value from the uniform distribution over the interval that starts with time of treatment discontinuation (in study days) and ends at the end of study.

The time to first moderate or severe COPD exacerbation will be analyzed up through Week 52 using a Cox regression model. Treatment comparisons will be performed using the model which includes treatment, baseline post-bronchodilator percent predicted FEV_1, baseline blood eosinophil count, baseline COPD exacerbation history ($1, \geq 2$), country, and ICS use at Screening (Yes/No). SAS PROC PHREG will be used. Estimated adjusted hazard ratios relative to the comparator for each treatment comparison will be displayed along with the associated Wald two-sided 95% confidence intervals and p-values (*Tables 2.8.1 to 2.8.3* for the efficacy, attributable, and treatment policy estimands, respectively).

Time to first moderate or severe COPD exacerbation will be displayed graphically for each treatment using a Kaplan-Meier curve and analyzed using a log-rank test to compare the curves between the treatments (*Figures 2.8.1 to 2.8.3* for the efficacy, attributable, and treatment policy estimands, respectively, and *Listing 6.1.2.3*). Subjects who complete the study on treatment and do not experience a COPD exacerbation over the treatment period will be censored at the date of study completion. For the efficacy estimand, subjects who discontinue study treatment early without experiencing a COPD exacerbation will be censored at the date of treatment discontinuation. For the treatment policy estimand, subjects who complete or withdraw from the study without experiencing a COPD exacerbation will be censored at the date of the last assessment or contact (including telephone contact).

Kaplan-Meier curves for time to moderate or severe COPD exacerbations will be constructed for completers vs. discontinuations (*Figure 2.8.1.1*), and, in separate plots, for MAR/MCAR vs. MNAR, and attributable vs. non-attributable discontinuation types (*Figures 2.8.1.2 and 2.8.1.3*).

Tipping Point Analyses for Time to First Moderate or Severe COPD Exacerbation

Robustness of results to missing data will be explored using tipping point analyses (Ratitch 2013). A brief overview of the approach is summarized in the table below.

Table 5 Sensitivity Analyses for Time to First Moderate or Severe COPD Exacerbations

Efficacy Estimand		Attributable Estimand	Treatment Policy Estimand
mITT Population		mITT Population	ITT Population
Tipping point analysis #1: All missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values that are considered MNAR are imputed with the rate in the treatment arm increased by up to 1.5 exacerbations/year until the p-value ≥ 0.05.	Tipping point analysis #2: All missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values are imputed with the rate in the treatment arm increased by up to 1.5 exacerbations/year until the p-value ≥ 0.05.	Tipping point analysis: MI using the 95th percentile of the reference arms' distribution (for the rate of moderate or severe COPD exacerbation) if treatment discontinuation is due to tolerability or lack of efficacy of study drug (as in the primary analysis of this estimand). Otherwise all missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values are imputed with the rate in the treatment arm increased by up to 1.5 exacerbations/year until the p-value ≥ 0.05.	Tipping point analysis: All missing data are imputed using the observed data model except that for subjects in the treatment arm (not the comparator arm) values are imputed with the rate in the treatment arm increased by up to 1.5 exacerbations/year until the p-value ≥ 0.05.

MNAR = Missing not at random.

The multiple imputation will be applied to the moderate or severe COPD exacerbation events within the negative-binomial analysis framework for the rate of moderate or severe COPD exacerbations, using values of δ that increase the rate in the treatment arm by up to 1.5 exacerbations/year. For this method, an underlying negative binomial stochastic process for the rate of exacerbations is assumed and post-treatment-discontinuation counts will be imputed conditional upon the reason for treatment discontinuation (see Appendix 6 for details).

A dataset with event counts through end of study will be created; however, the timing of the first imputed event is also needed. These times will be obtained for each imputed event by randomly drawing a value from the uniform distribution over the interval that starts with time of treatment discontinuation (in study days) and ends at the end of study. Missing values will first be imputed for the missing COPD exacerbation events prior to the computation of the time to the first

moderate or severe COPD exacerbation (for the sensitivity analysis). After imputation, the analysis will proceed to use Cox regression (as described above) and subsequently the multiple-imputation results will be combined using Rubin's formulae [Rubin, 1987].

6.4.5.2 Time to First Clinically Important Deterioration

Clinically important deterioration (CID) is defined as ≥ 100 mL decrease from baseline in trough FEV_1, or ≥ 4 points increase from baseline in SGRQ total score, or a TDI focal score of -1 point or less, or treatment-emergent moderate-or-severe COPD exacerbation occurring up to Week 52. Time to CID is a secondary endpoint in the Ex-US approach.

Time to first CID analysis will be performed using a Cox regression model. The model will include treatment, baseline post-bronchodilator percent predicted FEV_1, baseline blood eosinophil count, baseline COPD exacerbation history (1, ≥ 2), country, and ICS use at Screening (yes/no)**.** Time to a CID event will be based on the component event which occurs first. Subjects who do not experience a CID event will be censored at the earliest day among the component censoring times. COPD exacerbations happening after Week 52 will not be counted as CID events. Estimated adjusted hazard ratios will be displayed along with associated 95% CI and p-values (*Tables 2.11.1 to 2.11.3* for the efficacy, attributable, and treatment policy estimands, respectively). Time to first CID will be displayed graphically for each treatment group using a Kaplan-Meier curve and analyzed using a log-rank test to compare the curves between the treatments (*Figures 2.11.1 and 2.11.3* for the efficacy and treatment policy estimands, respectively).

For the analysis of the attributable estimand, missing data that have been reasonably attributed to tolerability or lack of efficacy will be imputed based on either the 5^{th} or the 95^{th} percentile of the reference arm's (FF MDI) distribution. The attributable estimand for time to CID will be computed by applying the percentile penalty to each of the four component variables simultaneously; it uses multiple by-visit imputation of pre-dose trough FEV_1 (at visits for which they are missing) (as described in Sections 6.4.4.3), of the SGRQ total score (as described in Section 6.4.5.4), of the TDI focal score in a manner similar to that for SGRQ, and MI of time to first moderate-or-severe COPD exacerbation (Section 6.4.5.1). A complete dataset for the COPD exacerbation count will be created analogously to that described in the Details Appendix to this SAP; however, the timing of imputed events is also needed. These will be obtained for each imputed event by randomly drawing a value from the uniform distribution over the interval that starts with time of treatment discontinuation (in study days) and ends at the end of study.

6.4.5.3 Transition Dyspnea Index

TDI focal score over 24 weeks is a secondary efficacy endpoint in the Ex-US approach, and an 'other endpoint' over 12 weeks, over 52 weeks, and at each post-randomization visit. Individual components of the TDI (functional impairment, magnitude of task, and magnitude of effort) are also 'other endpoints' over 12, 24 and 52 weeks, and at each post-randomization visit.

Assessments of dyspnea will be obtained using the BDI/TDI (where BDI is the Baseline Dyspnea Index). The BDI/TDI questionnaire can be found in Protocol Appendix 6.

At Randomization (Visit 3), the severity of dyspnea at baseline will be assessed using the BDI. BDI components are functional impairment, magnitude of task, and magnitude of effort (*Listing 6.1.6)*. The possible range of values for each BDI component score is 0 (very severe impairment) to 4 (no impairment). The BDI component scores are summed to determine the BDI focal score (0 to 12) (i.e., the lower the score, the worse the severity of dyspnea).

At subsequent visits (as per Schedule of Events: see of the Schedule of Events in the protocol), change from baseline will be assessed using the TDI. TDI components include: Change in Functional Impairment, Change in Magnitude of Task, and Change in Magnitude of Effort (*Listing 6.1.6)*. The TDI component score ranges from -3 (major deterioration) to +3 (major improvement). The sum of all component scores yields the TDI focal score (-9 to +9) (i.e., the lower the score, the more deterioration from baseline).

The difference between treatment groups in TDI focal score over 12, 24, and 52 weeks and at each post-randomization visit will be analyzed using a similar RM approach as for the primary endpoint, but using BDI instead of baseline FEV_1 in the model, and adding baseline post-bronchodilator percent predicted FEV_1 as a continuous covariate. Thus the model will include treatment, visit, treatment by visit interaction, and ICS use at Screening as categorical covariates and BDI, baseline blood eosinophil count, percent reversibility to Ventolin HFA and baseline post-bronchodilator percent predicted FEV_1 as continuous covariates. Scoring and handling of missing items will be conducted in accordance with the user's guide for the TDI score. Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference (*Table* and *Figure 2.2.1* for the efficacy estimand). Supportive analyses for the attributable estimand and treatment policy estimand will be conducted (*Tables* and *Figures 2.2.2* to *2.2.3*, respectively).

The attributable estimand will be computed in a similar manner as the attributable estimand is computed for change from baseline in morning pre-dose trough FEV_1 at Week 12 as described in Section 6.4.4.1.

In addition, the difference between treatments for the individual components of the TDI: functional impairment, magnitude of task, and magnitude of effort will each be analyzed over 12, 24, and 52 weeks, and at each post-baseline visit for the efficacy estimand using the same modeling approach as for the TDI focal score *(Table 2.2.4* and *Figures 2.2.4.1-2.2.4.3* for the functional impairment, the magnitude of task and the magnitude of effort for the efficacy estimand).

Furthermore, responder analyses will be performed for the TDI focal score where responders are defined as subjects with a response of 1.0 points or more (corresponding to at least a minor improvement) on average over 12, 24, and 52 weeks. Logistic regression will be used to compare the treatment groups with BDI, baseline blood eosinophil count, and percent reversibility to Ventolin HFA as continuous covariates and treatment, and ICS use at Screening as categorical covariates. P-values and odds ratios with 95% CIs will be produced for each treatment comparison *(Table 2.2.5* for the efficacy estimand).

For the TDI, at each visit, if a response to any of the three questions is missing, then the focal score will also be considered missing. For the TDI responder analyses, subjects without post-baseline data will be considered to be non-responders in the analysis.

TDI and BDI data will be listed in *Listing 6.1.6*.

6.4.5.4 St. George's Respiratory Questionnaire

Percentage of subjects achieving an MCID of ≥4 in SGRQ total score is a secondary endpoint in the US approach at Week 12 and in the Ex-US approach over 24 Weeks. It is also an 'other point' over 52 weeks and at each post-randomization in-clinic visit. The change from baseline in SGRQ total score and in individual domains of the SGRQ (Symptoms, Activity, and Impacts) over 12, 24 and 52 weeks, and at each post-randomization in-clinic visit, are 'other endpoints'.

The SGRQ will be used to provide the health status/health-related QoL measurements in this study (see Protocol Appendix 7). The SGRQ contains 50 rated items divided into three domains: "Symptoms" concerned with respiratory symptoms, their frequency, and severity; "Activity" concerned with activities that cause or are limited by breathlessness; and "Impacts" which covers a range of aspects concerned with social functioning and psychological disturbances resulting from airway disease. Individual items of SGRQ data will be listed (*Listings 6.1.7* for All Subjects Randomized).

A score will be calculated for each component and a "Total" score will also be calculated (*Listings 6.1.9* for All Subjects Randomized). In each case, the lowest possible value is zero and the highest is 100. Higher values correspond to greater impairment of QoL.

Scoring and handling of missing items will be conducted in accordance with the user's guide for the SGRQ. Each response is to be given a unique empirically derived weight between 0 and 100; the weights of all responses are then summed up and divided by the maximum possible score and expressed as a percentage.

Responder analyses of SGRQ as a secondary efficacy endpoint will be performed, where responders are defined as subjects with an improvement of ≥4.0 points at week 12, on average over 24 weeks, over 52 weeks, and at each post-randomization in-clinic visit. For the SGRQ responder analyses, subjects without post-baseline data or who discontinue treatment for any reason will be considered to be non-responders for the analysis. Logistic regression will be used to compare the treatment groups with baseline SGRQ score, baseline blood eosinophil count, baseline post-bronchodilator percent predicted FEV_1, and percent reversibility to Ventolin HFA as continuous covariates, and treatment and ICS use at Screening as categorical covariates. P-values and odds ratios with 95% CIs will be produced for each treatment comparison (*Tables 2.3.3 to 2.3.5* for the efficacy, attributable, and treatment policy estimands, respectively).

The attributable estimand for responder analysis of SGRQ will be computed as follows. First, multiple imputations will be performed on the continuous total SGRQ scores in a similar manner as for the attributable estimand that is computed for change from baseline in morning pre-dose trough FEV_1 at Week 12 as described in Section 6.4.4.1 except that the 95th percentile will be

used instead of the 5th percentile. After that, it will be determined whether the subject has attained the MCID. The analysis will proceed using logistic regression as described above, followed by combining of results across the multiple imputations using the formulae of Rubin [Rubin, 1987].

The difference between treatment groups in the change from baseline in SGRQ over 12 weeks, 24 weeks, over 52 weeks, and each post-baseline visit will be evaluated using a similar RM approach as for TDI focal score. The model will include treatment, visit, and treatment by visit interaction, ICS use at Screening as categorical covariates and baseline SGRQ score, baseline blood eosinophil count, percent reversibility to Ventolin HFA, and baseline post-bronchodilator percent predicted FEV_1 as continuous covariates. Missing SGRQ total scores will not be imputed. Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference (*Table and Figure 2.3.1* for the efficacy estimand).

Individual domains of the SGRQ will also be analyzed in a similar fashion as the total score (*Table 2.3.2* for the efficacy estimand).

Tipping Point Analyses for Percentage of Subjects achieving an MCID of 4 Units or More in SGRQ Total Score at Week 12

Robustness of results to missing data will be explored using tipping point analyses (Ratitch 2013). Details of the methods are similar to sensitivity analyses of FEV_1 (found in Section 6.4.4.3 and in the Statistical Methods Appendix to this SAP), but with the model for SGRQ total score described above and with the maximum increase 16 units for change from baseline in SGRQ total score for the tipping point analyses.

6.4.5.5 Rescue Ventolin HFA Use

Change from baseline in average daily rescue Ventolin HFA use is a secondary endpoint in the US approach over 12 Weeks and in the Ex-US approach over 24 Weeks. It is also an 'other point' for over 52 weeks and over each 4-week interval of the 52-week treatment period. Percentage of days with 'no rescue Ventolin HFA use' is an 'other endpoint'.

The number of puffs of rescue Ventolin HFA taken in the previous 12 hours since the previous (AM or PM) dose will be recorded in the subject diary in the morning and evening. The mean daily number of puffs of rescue Ventolin HFA used by subjects during the study will be calculated overall and for each of the 4-week intervals of the 52-week treatment period and provided in a diary data listing (*Listing 6.1.3* for All Subjects Randomized). Diary data recorded during the last 7 days of the Screening Period will be used to calculate the baseline.

For every interval of time over which the mean number of puffs of rescue will be calculated, records with missing values will be ignored in both the numerator and denominator. As such, the denominator will be adjusted based on the number of days (including half days) with non-missing values. That is, the mean daily number of puffs of daytime rescue use (M_DT) will be set to the total number of daytime puffs divided by the number of half-days when daytime rescue use was recorded. The mean daily number of puffs of nighttime rescue use (M_DN) will

be set to the total number of nighttime puffs divided by the number of half-days when the nighttime rescue use was recorded. The mean daily rescue use (puffs) is then two multiplied by the mean of M_DT and M_DN.

The difference between treatment groups in the change from baseline in rescue Ventolin HFA usage over 12, 24 and 52 weeks will be evaluated using a linear RM analysis of covariance (ANCOVA) model which will include treatment, 4-week time interval (1 to 13), treatment by 4-week interval interaction, and ICS use at Screening as categorical covariates, and baseline rescue Ventolin HFA use, baseline blood eosinophil count, baseline FEV_1, and percent reversibility to Ventolin HFA as continuous covariates. A UN matrix will be used to model additional autocorrelation within subject. If this model fails to converge, then an AR(1) structure will be used instead. In the AR(1) model, subject will be included as a random effect.

Contrasts will be used to obtain estimates of the treatment differences over the treatment period. Two-sided p-values and point estimates with two-sided 95% CIs will be produced for each treatment difference. The main analysis will be conducted using the efficacy estimand (*Table* and *Figure 2.4.1*). Supportive analyses for the attributable estimand, treatment policy estimand, and RVU population will be performed (*Tables* and *Figures 2.4.2, 2.4.3 and 2.4.6*).

The attributable estimand (for the analysis of average daily rescue Ventolin HFA use) will be computed in a similar manner as the attributable estimand is computed for change from baseline in morning pre-dose trough FEV_1 at Week 12 as described in Section 6.4.4.1 except that the 95^{th} percentile will be used instead of the 5^{th} percentile.

As supportive analyses, the treatment difference for each 4-week interval will be evaluated and summarized. Additionally, as supportive analyses, daytime rescue Ventolin HFA use and night-time rescue Ventolin HFA use will be evaluated and summarized in a similar fashion. Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference *(Tables* and *Figures 2.4.4 and 2.4.7* for the daytime rescue Ventolin use, and *2.4.5 and 2.4.8* for the nighttime rescue Ventolin use, for the efficacy estimand and RVU population).

Tipping Point Analyses for Rescue Ventolin HFA Use Over 12 Weeks

Robustness of results to missing data will be explored using tipping point analyses (Ratitch 2013). Details of the methods are similar to sensitivity analyses of FEV_1 (found in Sections 6.4.4.3 and in the Details Appendix to this SAP, and using the maximum value of $\delta = 4$ puffs), but with the model for rescue Ventolin HFA use described above. Multiple-imputation results will be combined using Rubin's formulae [Rubin, 1987].

Additional sensitivity analyses will be implemented based on a cumulative responder approach as described in Farrar 2006 for the change from baseline in average daily rescue medication over 12 weeks (*Tables 2.4.10.1 and 2.4.10.2* for the efficacy and treatment policy estimands, respectively). A cumulative distribution plot by treatment arm (Farrar et al., 2006) will also be produced. The observed change from baseline in average daily rescue medication over 24 weeks will be plotted on the X axis, while the proportion of responders (subjects that equal or exceed

that level of change) will be plotted on the Y axis *(Figure 2.4.10.1 and 2.4.10.2* for the efficacy and treatment policy estimands, respectively). Subjects without post-baseline data or who discontinue treatment for any reason will be considered non responders for all values. For display purposes only, the range of the x-axis will be from the 1st percentile to the 99th percentile irrespective of treatment in order to avoid the undue influence of outlying values. The cumulative responder curves for each treatment will then be compared pairwise using Kolmogorov-Smirnov tests. Cumulative responder analysis for the attributable estimand will not be performed as methodology to apply Rubin's rules for combining multiply imputed data for such an analysis is not readily available.

The percentage of days with 'no rescue Ventolin HFA use' over 12, 24, and 52 weeks, and for each 4-week interval of the Treatment Period will be evaluated with an ANOVA model. A 'day with no rescue use' is defined as any day where the subject reported zero puffs of rescue Ventolin HFA. The rescue Ventolin HFA usage diary data from days where rescue Ventolin HFA usage data is non-missing will be used to ascertain the days with "no rescue Ventolin HFA use". The percentage of days with no rescue use will be calculated as 100 x (number of days no rescue Ventolin use over the entire treatment period / number of days with non-missing rescue Ventolin use over the entire treatment period). For the efficacy estimand, days after discontinuation of study medication will not be used. The ANCOVA model will evaluate treatment differences and include baseline rescue Ventolin HFA use and percent reversibility to Ventolin HFA as continuous covariates and region, smoking status at baseline, and ICS use at Screening as categorical covariates. Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference *(Tables 2.5.1* to *2.5.2* for the efficacy estimand and RVU population, respectively).

6.4.5.6 EXACT Total Score

Change from baseline in EXACT total score over 24 weeks is a secondary efficacy endpoint in the Ex-US approach. The 11-item RS-Total Score, as well as the 3 symptom domain scores for breathlessness, cough and sputum, and chest symptoms over 24 weeks and 52 weeks, and over each 4-week interval of the 52-week treatment period are 'other endpoints'.

The EXACT is a 14-item patient reported outcome (PRO) from the daily diary which will be used to measure the effect of treatment on exacerbations, and on the severity of respiratory symptoms. Mean change from baseline in the daily EXACT Total Score, the 11-item E-RS Total Score, as well as 3 subscale scores, RS-Breathlessness, RS-Cough and Sputum, and RS-Chest Symptoms, will be calculated over each post-randomization 4-week interval of the treatment period. The last 7 days of the Screening Period will be used to calculate the baseline.

The mean change from baseline in the EXACT Total Score (*Tables* and *Figures 2.6.1* to *2.6.3* for the efficacy, attributable, and treatment policy estimands, respectively), E-RS Total Score (*Table* and *Figure 2.6.4* for the efficacy estimand), RS-Breathlessness, RS-Cough and Sputum, and RS-Chest Symptoms (*Table* and *Figure 2.6.5* for the efficacy estimand) over each 4-week interval will be analyzed using a similar RM model as for TDI to estimate treatment effects over 24 and 52 weeks, but using the corresponding baseline mean score instead of the BDI as a covariate.

Instead of visit, the number of the relevant respective 4-week interval (Interval 1 to Interval 13) will be used as a categorical covariate in the model. Thus the model will include treatment, time interval, and treatment by time-interval interaction, and ICS use at Screening as categorical covariates and baseline score, baseline blood eosinophil count, percent reversibility to Ventolin HFA, and baseline post-bronchodilator percent predicted FEV_1 as continuous covariates. A UN correlation matrix will be used to model additional autocorrelation within subject. If this model fails to converge, then an AR(1) structure will be used instead. In the AR(1) model, subject will be included as a random effect.

Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference. EXACT data will be listed in *Listing 6.1.5*.

The attributable estimand (for the analysis of EXACT total score) will be computed in a similar manner as the attributable estimand is computed for change from baseline in morning pre-dose trough FEV_1 over 24 weeks as described in Section 6.4.4.1 except that the 95^{th} percentile will be used instead of the 5^{th} percentile.

6.4.6 Other Efficacy Analyses

6.4.6.1 Other Spirometry Endpoints

The analysis for between-treatment comparisons of changes from baseline in morning pre-dose trough FEV_1 over 12, 24, and 52 weeks, and at each post-randomization in-clinic visit has already been described in Section 6.4.4.1 (*Tables* and *Figures 2.1.1* to *2.1.3*).

Similar analyses will be conducted for forced vital capacity (FVC), peak expiratory flow rate (PEFR), and forced expiratory flow between 25% and 75% of FVC (FEF_{25-75}) over 12, 24 and 52 weeks, and at each post randomization in-clinic visit for the efficacy estimand. The baseline covariate for each model will be endpoint-specific (*Tables* and *Figures 2.7.1* to *2.7.*3 for change from baseline in morning pre-dose trough FVC, PEFR, and FEF_{25-75}).

6.4.6.2 Rate of COPD Exacerbations

<u>COPD Exacerbations</u>

A **COPD exacerbation** is defined as a change in the subject's usual COPD symptoms that lasts two or more days, is beyond normal day-to-day variation, is acute in onset, and may warrant a change in regular medication. The change in symptoms must include at least one major COPD symptom and at least one other major (dyspnea, sputum volume, and sputum color) or minor symptom (cough, wheeze, sore throat, cold symptoms such as rhinorrhea or nasal congestion, and fever without other cause).

Exacerbations will be considered **moderate** if they result in:

- Use of systemic corticosteroids and/or antibiotics for at least three days; a single depot injectable dose of corticosteroids will be considered equivalent to a 3-day course of systemic corticosteroids.

Exacerbations will be considered **severe** if they result in:

- An inpatient COPD-related hospitalization (documentation stating that the subject was hospitalized for the COPD exacerbation or a record of the subject being admitted for ≥24 hours to an observation area, the emergency department, or other equivalent healthcare facility depending on the country and healthcare system).

- COPD-related death.

Moderate and severe COPD exacerbations will be entered in the eCRF.

Additionally, the Investigator may identify certain events (recorded on the same CRF page) which don't entirely meet the criteria above as exacerbations; the justifications supporting the Investigator's judgment will be recorded on a separate page on the eCRF.

COPD exacerbations not meeting the criteria for moderate or severe COPD exacerbations will be considered to be mild COPD exacerbations. For more detail about moderate-or-severe, severe, and any-severity COPD exacerbation events (and their start and end dates) and how they are operationally defined, see the subsections titled "Duration of COPD Exacerbation," "Moderate-or-Severe Exacerbation and Severe Exacerbation: Operational Definitions", and "Exacerbation of any Severity: Operational Definition".

The rate of COPD exacerbations of any severity, moderate or severe COPD exacerbations, severe COPD exacerbations, COPD exacerbations treated with systemic steroids, and COPD exacerbations treated with antibiotics will be analyzed using negative binomial regression as implemented in SAS PROC GENMOD. Treatments will be compared adjusting for baseline post-bronchodilator percent predicted FEV_1 and baseline blood eosinophil count as continuous covariates and baseline COPD exacerbation history (1, ≥ 2), and ICS use at Screening (yes/no) as categorical covariates. COPD exacerbations will be considered separate events provided that there are more than 7 days between the recorded stop date of the earlier event and the start date of the later event. Time at risk of experiencing an exacerbation will be used as an offset variable in the model.

For the efficacy estimand, the time at risk is defined as time of exposure to randomized treatment – not during or within 7 days after an exacerbation of equal or greater severity – until the last dosing date – not during or within 7 days after an exacerbation of equal or greater severity – up to the last recorded date of any assessment or contact for the subject (including telephone contact).

For moderate-or-severe COPD exacerbations that were identified apart from an eDiary alert, the symptom information is listed in *Listing 6.1.2.2*.

The number of exacerbations, the percentage of subjects who experience exacerbations, exacerbation rates, and rate ratios comparing treatments will be summarized for the efficacy estimand for moderate-or-severe exacerbations (*Table 2.8.9)*, for any severity of exacerbations

(*Table 2.8.10)*, for severe COPD exacerbations (*Table 2.8.11)*, for COPD exacerbations treated with systemic steroids (*Table 2.8.12*), and for COPD exacerbations treated with antibiotics (*Table 2.8.13*).

Follow-up time at risk will be summarized and displayed graphically for moderate or severe COPD exacerbations for the efficacy estimand (*Table and Figure 2.8.9a*), for any severity of exacerbations (*Table and Figure 2.8.10a)*, for severe COPD exacerbations (*Table 2.8.11a*), for COPD exacerbations treated with systemic steroids (*Table 2.8.12a*), and for COPD exacerbations treated with antibiotics (*Table 2.8.13a*).

Duration of COPD Exacerbation

For moderate or severe exacerbations, the duration is defined by the length of prescribed treatment (using the eCRF COPD exacerbation page), whereas for mild exacerbations, the duration is defined by the length of symptoms.

For moderate or severe COPD exacerbations, the start date will be defined as the start date of prescribed treatment with a systemic corticosteroid or systemic antibiotic and the stop date will be defined as the latter of the last day of prescribed treatment with a systemic corticosteroid or systemic antibiotic (if applicable). If the subject dies before being treated or hospitalized, then the start and stop date will be the date of death. In order to ensure that the same event is not counted twice, consecutive or concurrent moderate or severe COPD exacerbations with equal to or fewer than 7 days between the recorded stop date of the earlier event and start date of the later event will be considered the same event and assigned the maximum severity between the two.

For mild COPD exacerbations, start date will be defined as the onset of worsened symptoms as recorded by the subject in the eDiary, and the stop date will be defined as the last day of worsened symptoms. In order to ensure that the same event is not counted twice, consecutive or concurrent mild COPD exacerbations with equal to or fewer than 7 days between the recorded stop date of the earlier event and start date of the later event will be considered the same event.

In addition, in order to not double-count exacerbations that are moderate or severe, eDiary data from dates within 7 days prior to or after a moderate or severe exacerbation will not be counted as additional mild COPD exacerbations. This implies that continuing worsened symptoms that meet the definition of a mild exacerbation would need to be present at least 2 days prior to the 7-day period immediately preceding the start date of a moderate or severe COPD exacerbation in order to be considered a separate event. Similarly, worsened symptoms would need to be present for at least 2 days after the 7-day period immediately following a moderate or severe COPD exacerbation to be considered a separate event.

Analyses of each severity of exacerbation will account for the time that subjects are at risk of having an exacerbation of that severity or greater. Time during or immediately following – i.e. within 7 days of – an exacerbation will not be considered as part of the time that the subject was at risk. However, time during or immediately following an exacerbation of lower severity will be included since, for example, a subject experiencing a mild exacerbation is still at risk of the event

increasing in severity and becoming a moderate exacerbation. Thus, for example, in the analysis of severe COPD exacerbations, subjects will still be considered to be at risk of an exacerbation even during or within 7 days after a mild or moderate exacerbation. Likewise, in the analysis of moderate-or-severe COPD exacerbations, subjects will still be considered to be at risk of an exacerbation even during or within 7 days after a mild exacerbation.

Moderate-or-Severe Exacerbation and Severe Exacerbation: Operational Definitions

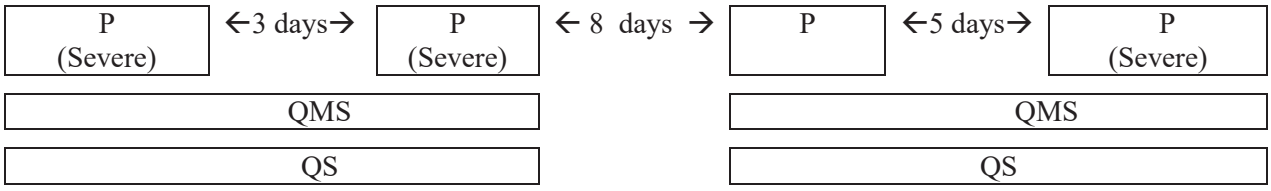
Moderate exacerbations and severe exacerbations will be defined based on information from the COPD Exacerbation eCRF page. A time interval from a single COPD exacerbation eCRF page will be designated as being during an event of a moderate-or-severe COPD exacerbation if either antibiotics or oral corticosteroids were administered for the exacerbation.

Call this time interval a "P-Interval". The start date of the P-Interval is the earliest start date of the above, and the stop date will be defined as the last stop day of the above. If the subject was hospitalized due to the exacerbation or if the exacerbation led to a COPD-related death, then the severity of "severe" will be assigned to this P-interval; otherwise the severity of "moderate" will be assigned. The later among the stop date of the treatment with a systemic corticosteroid and the stop date of the treatment with an antibiotic will be the end date of the COPD exacerbation (i.e. the end of the P-Interval).

An overarching interval of (any number of) such P-Intervals – including any P-Intervals with an end date not more than 7 days prior to the start date of some other P-Interval or with a start date not more than 7 days after the end date of some other P-Interval – and including the days in any gaps between them – will be called an "QMS-Interval". This QMS-interval will represent the consolidated duration of several exacerbations recorded on different CRF pages. This QMS-Interval will be considered to be a single event of a moderate-or-severe COPD exacerbation. See Figure 2.

A P-interval of severe COPD exacerbation is called a "severe" P-Interval. Any QMS interval that contains at least one "severe" P-Interval will also be called a "QS-Interval". This QS-Interval will be considered to be a single event of a severe COPD exacerbation. See Figure 2.

Figure 2 Overarching Intervals of Moderate-or-Severe (QMS) and Severe (QS) COPD Exacerbations



A P-interval is a moderate-or-severe COPD exacerbation instance from a single CRF page.
In a "Severe" P-Interval, the maximum severity of the COPD exacerbation is "severe".
A QMS interval is an overarching moderate-or-severe COPD exacerbation event encompassing multiple CRF pages.
A QS interval is an overarching severe COPD exacerbation event encompassing multiple CRF pages.

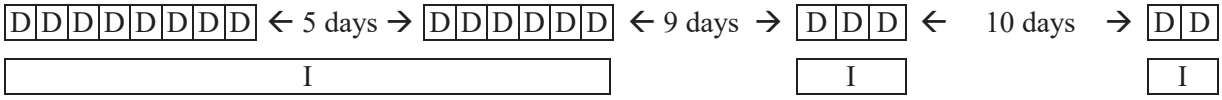
Exacerbation of any Severity: Operational Definition

Using eDiary data, a day will be designated as being during an event of a COPD exacerbation of some severity if (1) there was at least one major symptom and there was at least one other major or minor symptom and if (2) on an adjacent day there was at least one major symptom and there was at least one other major or minor symptom. Denote such a day as a "Category-D" day.

An interval of (any number of) such Category-D days – including any Category-D days not more than 7 days apart from some other Category-D day – and including the days in any gaps between them – will be called an "I-Interval". See Figure 3.

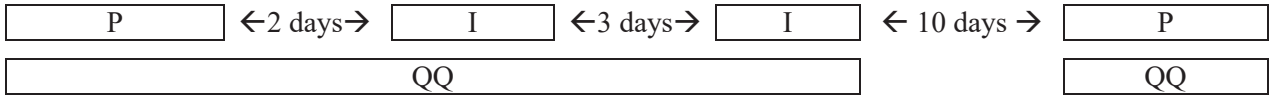
An overarching interval coalescing (any number of) P-Intervals and I-Intervals – including any such P-or-I-intervals with an end date not more than 7 days prior to the start date of some other P-or-I-Interval or with a start date not more than 7 days after the end date of some other P-or-I-interval – and including the days in any gaps between them – will be called a "QQ-Interval". This QQ-interval will represent the consolidated duration of several exacerbations recorded on different CRF pages or identified from subject diary data. This QQ-Interval will be considered to be a single event of an any-severity COPD exacerbation. See Figure 4.

Figure 3 Overarching Intervals (I) of Mild-Moderate-or-Severe COPD Exacerbation Events Based on e-Diary Symptom Data



A Category-D day is a day with mild-moderate-or-severe COPD exacerbation based on e-diary symptom data.
An I-Interval is an overarching mild-moderate-or-severe COPD exacerbation event encompassing multiple clusters of e-diary symptom days.

Figure 4 Overarching Intervals (QQ) of Mild-Moderate-or-Severe COPD Exacerbation Events Incorporating Both CRF Data and e-Diary Symptom Data



A P-Interval is a moderate-or-severe COPD exacerbation instance from a single CRF page.
An I-Interval is an overarching mild-moderate-or-severe COPD exacerbation event based on e-diary symptom data.
A QQ-Interval is an overarching mild-moderate-or-severe COPD exacerbation event – encompassing multiple P-Intervals and I-Intervals – incorporating both CRF data and e-diary symptom data.

In summary, we combine CRF based moderate-or-severe COPD exacerbation events (or severe COPD exacerbation events) if they are close enough together in time (Figure 2). We also combine mild-moderate-or-severe COPD exacerbations if they are close enough together in time;

this coalescing is done first within-data-source (CRF [Figure 2] or diary [Figure 3]) and then between the two sources (Figure 4).

Time-at-Risk for COPD Exacerbations of Various Severities: Operational Definition

During a time when a subject is not experiencing a severe COPD exacerbation (i.e. QS interval) – and is not in the seven days following a severe COPD exacerbation – a subject is considered to be at risk of having a severe exacerbation. During a time when a subject is not experiencing a moderate-or-severe COPD exacerbation (i.e. QMS interval) – and is not in the seven days following a moderate-or-severe COPD exacerbation – a subject is considered to be at risk of having a moderate-or-severe exacerbation. During a time when a subject is not experiencing an any-severity COPD exacerbation (i.e. QQ interval) – and is not in the seven days following an any-severity COPD exacerbation – a subject is considered to be at risk of having an any-severity exacerbation.

Overarching coalesced intervals (i.e. events) of COPD exacerbation will be listed for severe exacerbations, moderate -to-severe exacerbations, and any-severity exacerbations (*Listing 6.1.2.3*). A severe COPD exacerbation event must be classified also as a moderate-or-severe event and also as an any-severity event. A moderate-or-severe COPD exacerbation event must be classified also as an any-severity event.

Rate of COPD exacerbations of any severity will be analyzed in a manner similar to the rate of moderate or severe COPD exacerbations *(Table 2.8.10* for the efficacy estimand).

The count of COPD exacerbations of any severity is the number of QQ-Intervals (for a subject) as defined previously. Time at risk of experiencing an exacerbation will be used as an offset variable in the model. Time during an exacerbation (of any severity) or in the 7 days following an exacerbation (of any severity) will not be included in the calculation of exposure (i.e. time at risk). Data related to COPD exacerbations of any severity are listed in *Listings 6.1.2.1, 6.1.2.2, 6.1.2.3,* and *6.1.4*. For moderate-or-severe COPD exacerbations that were identified apart from an eDiary alert, the symptom information is listed in *Listing 6.1.2.2*.

6.4.6.3 Time to First COPD Exacerbation of any Severity

Time to first COPD exacerbation of any severity will be analyzed and graphically displayed using a similar approach as for time to first moderate or severe COPD exacerbations (*Table* and *Figure 2.8.5* for the efficacy estimand).

6.4.6.4 Time to First Severe COPD Exacerbation

Time to first severe COPD exacerbation will be analyzed and graphically displayed using a similar approach as for time to first moderate or severe COPD exacerbations (*Table* and *Figure 2.8.6* for the efficacy estimand).

6.4.6.5 Time to First COPD Exacerbation Treated with Systemic Steroids

Time to first COPD exacerbation treated with systemic steroids will be analyzed and graphically displayed using a similar approach as for time to first moderate or severe COPD exacerbations (*Table* and *Figure 2.8.7* for the efficacy estimand).

6.4.6.6 Time to First COPD Exacerbation Treated with Antibiotics

Time to first COPD exacerbation treated with antibiotics will be analyzed and graphically displayed using a similar approach as for time to first moderate or severe COPD exacerbations (*Table* and *Figure 2.8.8* for the efficacy estimand).

6.4.6.7 Time to Treatment Failure

Treatment failure is defined as a moderate or severe COPD exacerbation or discontinuation from treatment for any reason or death. Time to treatment failure will be displayed graphically for each treatment group using a Kaplan-Meier curve and analyzed using a log-rank test to compare the curves between the treatments (*Figure 2.9.1* for the efficacy estimand). Subjects who do not experience a treatment failure will be censored at Day 365 or the End of Study day whichever comes first. The time to treatment failure will be analyzed using the efficacy estimand. The model will include treatment, baseline post-bronchodilator percent predicted FEV_1, baseline blood eosinophil count, baseline COPD exacerbation history ($1, \geq 2$), and ICS use at Screening (yes/no). Estimated adjusted hazard ratios will be displayed along with associated 95% CI and p-values (*Table 2.9.1* for the efficacy estimand).

6.4.6.8 Time to Death: All Causes, Respiratory

The primary analysis will use the treatment policy estimand. Time to death (all causes) in weeks will be summarized using a Kaplan-Meier curve. Subjects who do not die will be censored at the day of last contact or confirmed vital status. A Cox regression model will be used to compare the treatments and include treatment, baseline post-bronchodilator percent predicted FEV_1, and baseline age. Estimated adjusted hazard ratios will be displayed along with associated 95% CI and p-values (*Table* and *Figure 2.10.1* for the treatment policy estimand).

Data permitting, these analyses will be repeated for time to death from respiratory causes. Only those deaths identified as being due to respiratory causes by the Clinical Endpoint Committee will be considered as events for this analysis. In addition to the censoring rules above, subjects who die from non-respiratory causes will be censored at the day of death. Estimated adjusted hazard ratios will be displayed along with associated 95% CI and p-values (*Table* and *Figure 2.10.2* for the treatment policy estimand).

The analysis of time to death will be conducted contingent upon having at least 30 events of death. Otherwise, the analysis will be limited to counts and listings.

6.4.6.9 Time to First Sustained Clinically Important Deterioration

Sustained CID is defined as $\geq$ 100 mL decrease from baseline in trough FEV_1, or $\geq$ 4 points increase from baseline in SGRQ total score, or a TDI focal score of -1 point or less, any of which

has occurred on two consecutive analysis visits or for ≥ 50% of all available subsequent analysis visits, or a treatment emergent moderate-or-severe COPD exacerbation occurring up to Week 52. Time to sustained CID is an "other" endpoint.

The number of subsequent scheduled visits is determined based on the remaining visits that a given patient is expected to complete before End of Study occurs (see section 6.4.3 for the planned last visit definition).

Time to sustained CID will be analyzed by a Cox model, which will include treatment, baseline post-bronchodilator percent predicted FEV_1, baseline blood eosinophil count, baseline COPD exacerbation history (1, ≥ 2), and ICS use at Screening (yes/no)**.** Time to a sustained CID event will be based on the component event which occurs first. Subjects who do not experience a sustained CID event will be censored at the earliest day among the component censoring times. COPD exacerbations happening after Week 52 will not be counted as sustained CID events. Estimated adjusted hazard ratios will be displayed along with associated 95% CI and p-values (*Tables 2.11.4* for the efficacy estimand). Time to a sustained CID will be displayed graphically for each treatment group using a Kaplan-Meier curve and analyzed using a log-rank test to compare the curves between the treatments (*Figure 2.11.4* for the efficacy estimand).

Subjects who do not experience a sustained CID exacerbation and complete the study will be censored at the date of study completion. For the efficacy estimand, subjects who discontinue study treatment early without experiencing a sustained CID will be censored at the date of last study drug dose.

6.4.6.10 European Quality-of-Life-5 Dimension-5 Level Questionnaire (EQ-5D-5L)

The European Quality-of-Life-5 Dimension-5 Level Questionnaire (EQ-5D-5L) data will be weighted to calculate an index score based upon subjects' responses to the 5 dimensions. The visual analogue scale (VAS) will be scored from 0 (worst imaginable health state) through 100 (best imaginable health state) to represent the subject's self-report concerning how bad or how good their health was during that day.

EQ-5D will be presented in three different ways:
1. Presenting results from the EQ-5D-5L descriptive system as a health profile at baseline, at all visits, and at EoT (%, n) by domain
2. Presenting results of the VAS as a measure of overall self-rated health status - baseline scores, scores at each visit, changes from baseline at each visit, and mean VAS score over the treatment period
3. Presenting results from the EQ-5D-5L index score (using UK value set) baseline, each visit, changes from baseline to each visit, and the mean index score over the treatment period.

The percentage of subjects' categorical responses to each of the 5-dimensions will be summarized (*Table 2.12.1* for the efficacy estimand). Descriptive statistics for the index score (*Table 2.12.2* for the efficacy estimand) and VAS (*Table 2.12.3* for the efficacy estimand) will be presented by treatment group. VAS scores over 24 weeks may be analyzed using a similar

RM model as is used for the TDI, but using baseline EQ-5D VAS score as a covariate instead of BDI (*Table* and *Figure 2.12.2* for the index score and *Table 2.12.3* for VAS for the efficacy estimand). EQ-5D data are listed in *Listing 6.1.10*.

For calculations of index score, the method recommended by the national institute for health and care excellence (NICE) August 2017 will be applied. Cross-walk between EQ-5D-3L value set and EQ-5D-5L descriptive system have been developed by Van Hout et al 2012 (Van Hout *et al.* 2012) and this cross-walk value set for EQ-5D-5L will be used to calculate the index score (Van Reenan 2015). Appendix 8 contains the SAS/SPSS codes for crosswalk between 5L and 3L for calculation of index score.

No imputation will be made for missing data in either the EQ-5D-5L or VAS responses.

The compliance of completing the EQ-5D-5L questionnaires is a critical issue in the QoL and health-state evaluation, and will be described by post-randomization visit, by displaying the number and percentage of subjects who were assessed (per subject, at least 1 question answered) at each visit *(Table 2.12.4* for the efficacy estimand).

6.4.7 Subgroup Analyses

Subgroup analyses will be performed for change from baseline in morning pre-dose trough FEV_1, and rate of moderate or severe COPD exacerbations (efficacy estimand only). The following subgroups will be considered:

- History of Moderate or Severe COPD Exacerbation in the last 12 Months:
  - 1
  - ≥2
- Baseline Eosinophil Count:
  - <150 cells per mm^3
  - ≥150 cells per mm^3
- Country

Each subgroup will be analyzed separately using the same model that was used for the overall (combined subgroups) analysis. Estimates for the treatment effect and for the treatment differences will be displayed in the efficacy endpoint tables for each subgroup *(Tables 4.1.1* to *4.2.3).*

For each subgroup analysis, a test for the treatment-by-subgroup interaction will be performed using the same model that was used for the overall (combined subgroups) analysis but with the addition of terms for subgroup and the treatment-by-subgroup interaction. A table will be provided with the p-value for the test of the treatment-by subgroup interaction *(Table 4.3* for the efficacy estimand*)*.

Eosinophil Cut Point Exploration

Subgroup analyses of trough FEV_1 will be conducted in the baseline eosinophil count-high (≥150 cells per mm^3) and the baseline eosinophil count-low (<150 cells per mm^3) subgroups. It is acknowledged 150 cells per mm^3 may not ultimately be the appropriate threshold for evaluation of treatment benefit. Thus, additional analyses will evaluate alternative thresholds, and the results from these analyses could then inform thresholds for future clinical studies. This exploration will include using additive mixed models that combine nonparametric regression for the relationship of eosinophil levels to trough FEV_1 (*Figures 4.4.1, and 4.4.1.1 for trough FEV_1 at Week 12, and Figures 4.4.2, and 4.4.2.1 for trough FEV_1 Over 24 Weeks*) as well as potentially using subgroups defined by different cut points. A similar subgroup analysis will also be conducted for the rate of moderate or severe COPD exacerbations (*Figures 4.4.3 and 4.4.3.1).*

6.4.8 Control of Type I Error

The same general Type I error control strategy will be employed for the US and Ex-US regulatory approaches. In each approach, Type I error will be controlled for the following analyses:

- The primary analyses for the efficacy estimand
- The secondary analysis of the primary endpoints for the attributable estimand
- The secondary analyses
- A subgroup analysis of patients with 2 or more moderate or severe COPD exacerbations in the previous 12 months

6.4.8.1 US Approach

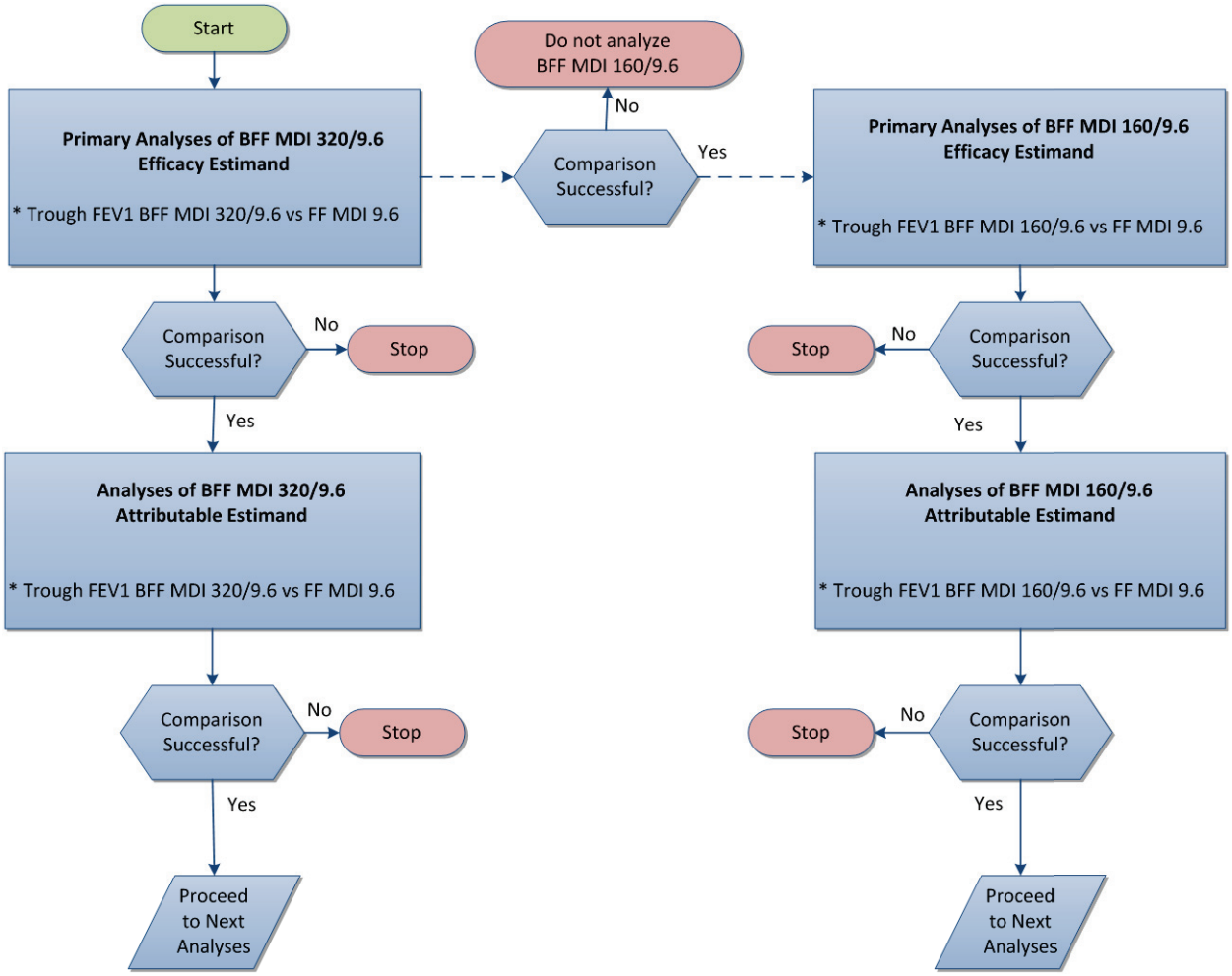
The Type I error rate will be controlled within the primary, secondary, and subgroup efficacy analyses. The primary analyses associated with BFF MDI 160/9.6 will proceed only if the primary analysis associated with BFF MDI 320/9.6 is successful. The analysis of the primary endpoint for the attributable estimand, secondary analyses, and subgroup analysis for each BFF dose will proceed only if the primary analysis associated with that dose of BFF MDI is successful.

For ease of review, the set of planned analyses has been divided into two groups:

Group 1: The primary analyses under the efficacy estimand and the analysis of the primary endpoints for the attributable estimand.

Group 2: The secondary and subgroup analyses under the efficacy estimand.

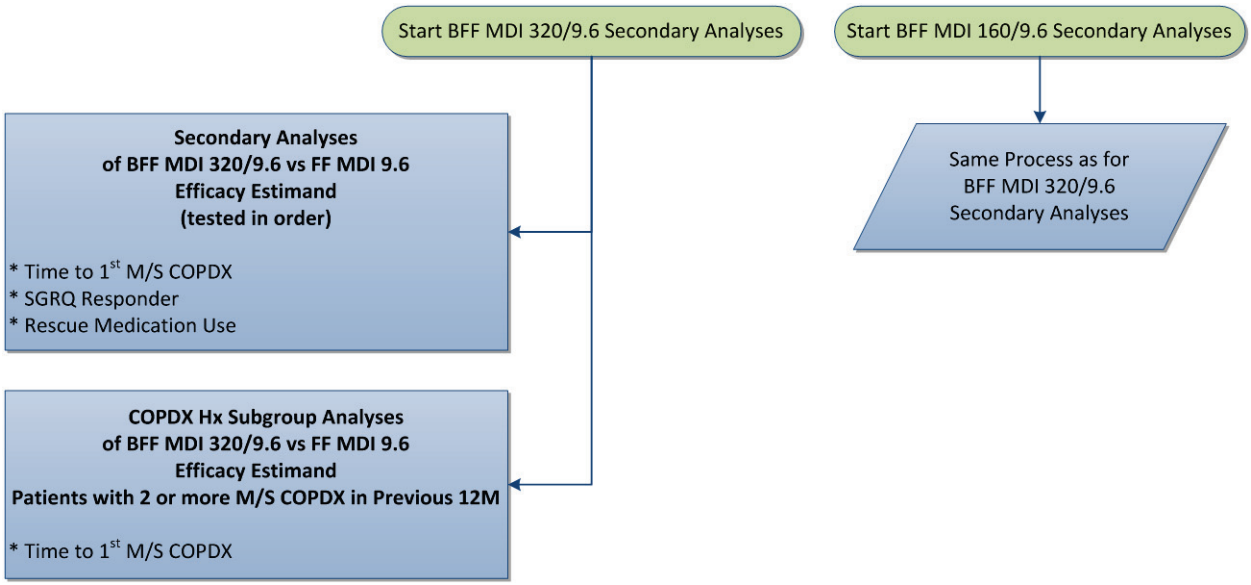
Figure 5 Group 1: Type I Error Control for the Analyses of the Primary Endpoints (US Approach)



In Group 1, a sequential multiplicity approach will be used in the analyses of the primary endpoint. In this approach, the analyses of the primary endpoint are listed in a pre-specified order in which they will be tested. Each hypothesis will be tested at the 2-sided 0.05 level. If a p-value is less than 0.05, then that hypothesis is rejected and the next hypothesis is tested. If the primary hypothesis is not rejected, then testing will stop. However, it is noted that p-values will still be calculated for all subsequent analyses for descriptive purposes.

Group 2 analyses are presented in Figure 6.

Figure 6 Group 2: The Analysis of the Secondary Endpoints (US Approach)

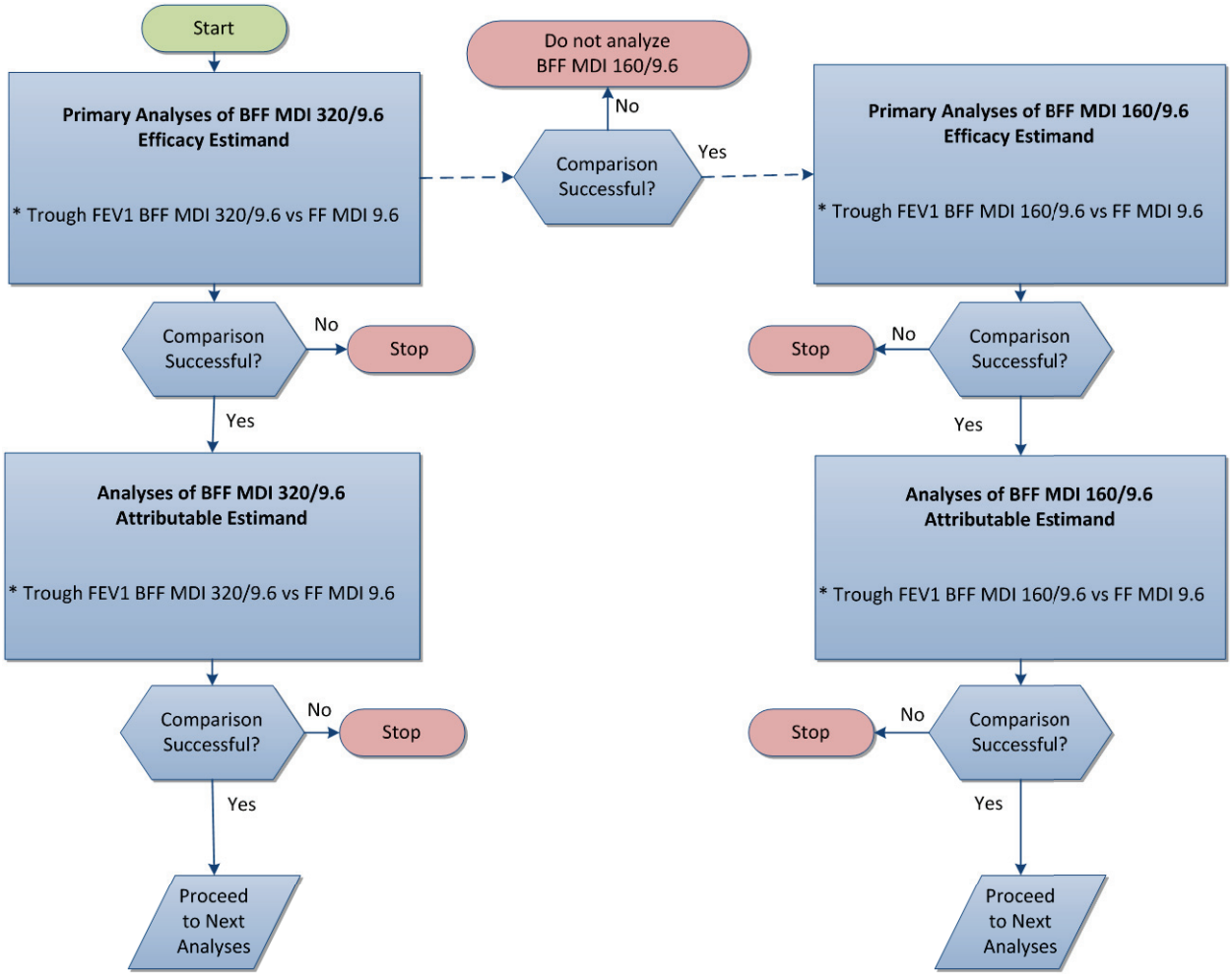


In Group 2, the secondary analyses and the subgroup analysis will be treated as separate families of hypotheses. The Type I error in each family of secondary analyses will be controlled to alpha.

6.4.8.2 Ex-US Approach

The Ex-US Approach will follow a similar general strategy as the US Approach. The control of Type I error in the Ex-US Approach differs from the US Approach in the number of secondary endpoints being evaluated. Otherwise, the Type I error control strategy is similar. The graphical representations of the primary, secondary, and subgroup analyses are presented below.

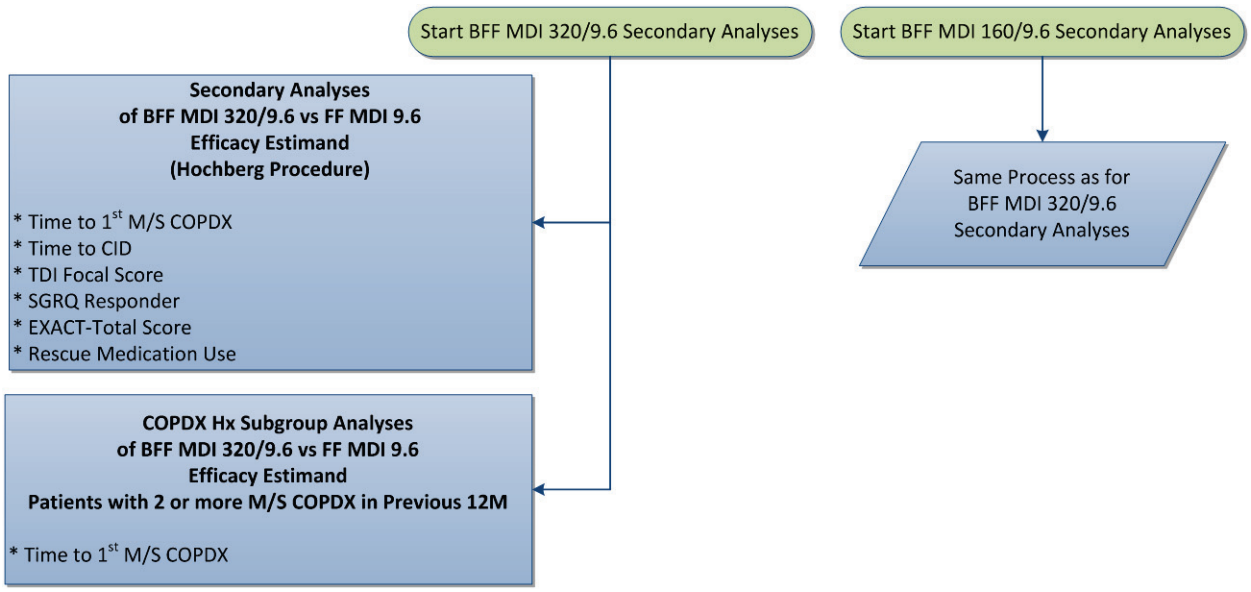
Figure 7 Group 1: Type I Error Control for the Analyses of the Primary Endpoints (Ex-US Approach)



If the Group 1 analyses are successful within a BFF MDI dose, the analysis of the secondary endpoints for that BFF MDI dose (Group 2) will proceed under the efficacy estimand.

Group 2 analyses are presented in Figure 8.

Figure 8 Group 2: The Analysis of the Secondary Endpoints (Ex-US Approach)



6.4.9 Correlation Analysis

Pearson correlation coefficients will be generated between the primary and secondary continuous endpoints (from both US and ex-US approaches). SGRQ total score will be used in place of the SQRQ responder. The mITT population will be used.

Note that for morning trough FEV_1, TDI, and change from baseline in SGRQ, the estimates over 24 weeks were obtained as LS means from MMRM analyses, and were not derived at the subject level. For the purpose of the correlation analysis, the endpoints over 24 weeks will be represented by simple averages of available data over 24 weeks.

The correlations will be organized in a matrix, with its upper triangle filled with pairwise Pearson correlation coefficients. All treatment groups will be pooled (Table 2.13.1), and also analyzed individually (Tables 2.13.2 to 2.13.4).

6.5 Safety Analysis

All safety analyses are based on the Safety Population. Hypothesis testing will not be performed for any safety analyses.

All AE data, clinically significant laboratory values, vital signs, and ECG values will be categorized according to their onset date into the following study periods:

- Events occurring during the treatment period (referred to as “On-Treatment” or “treatment-emergent”) are events with an onset date on or after the first date of dose and up to and

including the last day of randomized treatment (for study drug completers) or the last day of randomized treatment + 1 day (for premature treatment discontinuation). Events known to have occurred before the time of the first dose of study treatment are not included.

- Events occurring during the post-treatment-discontinuation follow-up are events with an onset date after the last day of randomized treatment (for study drug completers) or on or after the last day of randomized treatment + 2 days (for premature treatment discontinuation). The exception is that deaths are still considered to be during the Treatment Period if any adverse event that led to that death is during the Treatment Period.

Any AEs, clinically significant laboratory values, vital signs, and ECG values during the randomized-treatment period will be tabulated and listed. Beginning on the day after the date of discontinuation from or completion of study medication has passed, any new clinically significant ECGs, laboratory values, and vital signs will not be included in the tabulation or the computation of incidence rates, but will still be listed. Any new AEs, SAEs, and deaths during the post-randomized-treatment maintenance period will be tabulated and listed.

6.5.1 Adverse Events

The version of the Medical Dictionary for Regulatory Activities (MedDRA) that is current at the time of database lock will be used to code verbatim terms for AEs for final analysis of the data. A glossary of MedDRA preferred terms used for adverse events reported in the study along with the associated Investigator's verbatim will be provided in *Listing 7.2*.

An AE is considered treatment-emergent if an event occurs after the first dose of randomized study medication in the study, or if the AE worsened during the study after the first dose of study medication in the study (intensity and/or severity changed to a worsened grade), and the event onset is on or before the date of discontinuation from or completion of randomized study medication. Adverse events with onset date after the date of discontinuation from or completion of randomized treatment will not be considered treatment-emergent, but will be listed in adverse event data listings, and will be tabulated separately. Adverse events that occur between the time the subject signs the informed consent form for the study and the time when that subject is randomized are to be recorded as medical history unless the event met the definition of an SAE.

The incidence of an AE will be defined as the number and percentage of subjects experiencing an event. Adverse events will be tabulated at the level of the MedDRA preferred term and the MedDRA system organ class. No hypothesis tests will be performed.

An overview table will be prepared for the Safety Population with the incidences of subjects with at least one Treatment-Emergent Adverse Event (TEAE), at least one serious TEAE, at least one TEAE related to study treatment, at least one serious TEAE related to study treatment, at least one TEAE leading to premature treatment discontinuation, and a report of death (*Tables 3.1.1, 3.1.13, and 3.1.14*).

Events with Irregular Onset Dates

All TEAEs will be included in the data listings regardless of the completeness of the onset dates. Partial dates will be imputed in order to determine if an AE is treatment-emergent using the imputation rules in Appendix 1; however, imputed dates will not be provided in the data listings.

All adverse events, whether treatment-emergent or not, will be included in the listings. Reported adverse events by system organ class, preferred term, treatment, country, center, subject and onset day will be provided *(Listing 7.1)*. Reported adverse events by treatment, country, center, subject, and onset date will be presented in *Listing 7.3*. SAE-specific report information will be listed in *Table 3.8.2.1* and *Table 3.8.2.2*.

The listing of adverse events will provide the severity, maximum severity, relationship to study drug, action taken and outcome for each adverse event. Any SAEs reported will be listed for all subjects screened (*Tables 3.8.2.1).* Adverse events leading to permanent discontinuation of study treatment will be listed for the Safety Population *(Table 3.6.1)*. A listing of any reported deaths during the study (prior to randomization, during the randomized-treatment period, or during the post-randomized-treatment maintenance period) will be provided (*Table 3.15.2.1*); study treatment taken prior to the death and the number of days since the last dose of this study treatment at the time of the death will be included in the listing.

Summary tabulations of the following will be prepared for all subjects, for each treatment, for each primary system organ class, and for each preferred term within a system organ class:

- The incidence of all treatment-emergent adverse events *(Tables 3.2.1.1, 3.2.1.2, Table 3.2.1.3)*
- The incidence of subjects with adverse events by SOC during the post-randomized-treatment maintenance period (*Tables 3.2.1.18, 3.2.1.19)*
- The incidence of treatment-emergent adverse events occurring in SMQs (Standard MedDRA Queries)/groupings of interest *(Tables 3.2.3.1, 3.2.3.2)*
- The incidence of non-serious treatment-emergent adverse events occurring in ≥5% of subjects in a treatment *(Tables 3.2.4.1, 3.2.4.2)*
- The incidence of all treatment-related treatment-emergent adverse events *(Tables 3.4.1, 3.4.2)*
- The incidence of discontinuation from study treatment due to a treatment-emergent adverse events *(Tables 3.5.1, 3.5.2)*
- The incidence of treatment-emergent serious adverse events *(Tables 3.7.1.1.1, 3.7.1.1.2, 3.7.1.1.3)*
- The incidence of subjects with serious adverse events by SOC during the post-randomized treatment maintenance period (*(Tables 3.7.2.3, 3.7.2.4)*

- The incidence of all treatment-related treatment-emergent serious adverse events *(Tables 3.9.1, 3.9.2)*

- The incidence of all treatment-emergent adverse events by highest severity to treatment *(Tables 3.11.1.1* through *3.11.4.2* for the four treatments)

- The incidence of treatment-emergent adverse events occurring in at least 2% of subjects in any treatment *(Tables 3.2.2.1, 3.2.2.2* sorted by descending frequency of events in a preferred term).

- In addition, to control for possible differences in exposure between the treatments, the following AE and SAE summaries will be presented with the frequency and rate of occurrence (total number of events per 1000 person-years of exposure) by treatment, primary system organ class, and preferred term:

    - Frequency and rate of AEs *(Tables 3.3.1, 3.3.2)*
    - Frequency and rate of SAEs *(Tables 3.8.3.1, 3.8.3.2)*
    - Frequency and rate of neoplasms (*Tables 3.10.3.1* and *3.10.3.2* – All Cancer, *3.10.4.1* and *3.10.4.2* - Excluding Non-Melanoma Skin Cancer).

6.5.1.1 Adverse Events of Special Interest

Adverse events of special interest (AESIs) have been defined based on known effects of LABAs and ICS. These include but are not limited to cardiovascular, tremor effects, hyperglycemia, and hypokalemia for LABAs; and local (e.g., candidiasis and voice effects) and systemic (e.g., bone and skin effects, diabetes control, ocular and taste effects, adrenal suppression) steroid class effects and lung infection for ICS.

Standardized MedDRA queries (SMQs) will be utilized when possible, and a selection of high-level group terms (HLGTs), high-level terms (HLTs), and PTs will be utilized to represent other situations. The terms proposed to be used in the assessment of AESIs associated with ICS and LABAs are listed in Table 6. SMQs will be utilized when possible and a selection of preferred terms in other situations (*Appendix 5*).

Table 6 Adverse Events of Special Interest

Medical Concept	Selection of MedDRA Terms
Adrenal suppression	Adrenal cortical hypofunctions HLT
Agitation or anxiety	Collection of PTs
Bone fracture	Collection of HLGTs, HLTs, and PTs.
Candidiasis	Collection of PTs
Cardiovascular	Cardiac arrhythmias SMQ Cardiac failure SMQ Ischemic heart disease SMQ Torsades de Pointe/QT prolongation SMQ

Medical Concept	Selection of MedDRA Terms
Cardiovascular death	Collection of PTs
Cerebrovascular condition	CNS haemorrhages and cerebrovascular conditions SMQ
Diabetes mellitus	Hyperglycaemia/new onset diabetes mellitus SMQ
Dysgeusia or ageusia	Collection of PTs
Dysphonia or aphonia	Collection of PTs
Gastrointestinal	Gastrointestinal perforation, ulceration, haemorrhage or obstruction SMQ Gastrointestinal obstruction SMQ
Headache	Headache (PT)
Hypercortisolism	Collection of PTs
Hypertension	Blood pressure ambulatory increased (PT) Blood pressure increased (PT) Blood systolic increased (PT)
Hypokalemia	Collection of PTs
Lower respiratory tract infections other than pneumonia	Bronchitis (PT) Bronchitis viral (PT) Bronchitis bacterial (PT) Lower respiratory tract infection (PT) Lower respiratory tract infection viral (PT) Lower respiratory tract infection bacterial (PT) Infective exacerbation of chronic obstructive airway disease (PT)
Ocular effects	Visual disorders HLT Glaucoma SMQ increased intraocular pressure collection of PTs Cataract collection of PTs
Osteoporosis and osteopenia	Osteoporosis/osteopenia (SMQ)
Palpitation	Palpitations PT
Paradoxical bronchospasm	Collection of PTs
Pneumonia	Collection of PTs
Psychiatric effect	Collection of PTs
Skin effects	Skin atrophy (PT) Skin striae (PT) Acne (PT) Contusion (PT) Ecchymosis (PT) Increased tendency to bruise (PT) Petechiae (PT) Purpura (PT) Malassezia folliculitis (collection of PTs) Hypertrichosisa (collection of PTs) Alopecia (collection of PTs)

Medical Concept	Selection of MedDRA Terms
Sleep effects	Initial insomnia (PT) Insomnia (PT) Sleep disorder (PT)
Sudden death	Collection of PTs
Throat irritation	Collection of PTs
Tremor	Tremor HLT
Weight gain	Collection of PTs

Abbreviations: CNS=central nervous system; PT=preferred term.

Adverse Events in MedDRA SMQs/Groupings of Interest by Term will be tabulated *(Tables 3.12.1, 3.12.2)*.

6.5.1.2 Major Adverse Cardiovascular Event (MACE) Events Determined by Clinical Endpoint Committee

The clinical endpoint committee (CEC) will review and adjudicate serious Cardio- and Cerebrovascular (CCV) events as MACE. MACE events are defined as the following:

- Cardiovascular death
- Non-fatal myocardial infarction (MI)
- Non-fatal stroke

The CEC will review and assess these non-fatal serious CCV events and all deaths as to whether or not they fulfill criteria (based on CEC working practices) for MACE.

MACE events will be summarized by adjudicated CRF category and treatment group (*Tables 3.13.1.1, 3.13.1.2*). The assessment of MACE events will include the rate of confirmed MACE events (*Tables 3.13.2.1, 3.13.2.2*). Adjudicated MACE events will be listed in *Listing 7.4*.

The incidence of subjects with adjudicated MACE AEs by category will be summarized in *Table 3.13.3.1 and Table 3.13.3.2*.

6.5.1.3 Pneumonia Events Determined by Clinical Endpoint Committee

All AEs/SAEs with preferred terms that could relate to pneumonia will be adjudicated to provide a more complete assessment of all physician-reported pneumonias. The incidence of confirmed pneumonia events will be tabulated (*Tables 3.14.1.1, 3.14.1.2*). The assessment of pneumonia events will include the rate of confirmed pneumonia events (*Tables 3.14.2.1, 3.14.2.2*). Adjudicated pneumonia events will be listed in *Listing 7.4*.

In order to account for specific patient risk factors, data permitting, time to first pneumonia will be compared between treatments using Cox proportional hazards (*Tables 3.14.3.1, 3.14.3.2*).

Specific patient risk factors (baseline FEV_1, age, and medical history of pneumonia in the last 5 years [Yes or No]) will be evaluated for inclusion.

The incidence of subjects with adjudicated pneumonia AEs by category will be summarized in *Table 3.13.3.1 and Table 3.13.3.2*.

6.5.1.4 Cause of Death Determined by Clinical Endpoint Committees

Causes of death will be listed by subject and summarized by treatment for (1) all-cause mortality, (2) mortality of probable cardiovascular cause, (3) mortality of probable respiratory cause, (4) mortality of cancer and (5) mortality of probable other causes using the Safety Population based on (A) cases reported during the active Treatment Period and (B) cases reported during the active Treatment Period plus one day *(Tables 3.15.2.1 and 3.15.2.2)*. The incidence of subjects with a death event will be tabulated by adjudicated CRF category and treatment during the randomized-treatment period *(Tables 3.15.1.1, 3.15.1.2)* and during the post-randomized-treatment maintenance period (*Tables 3.15.1.3 and 3.15.1.4).* To control for possible differences in exposure between treatments, the death will be summarized with frequency and rate of occurrence (total number of events per 1000 person-years of exposure) by treatment, primary system organ class, and preferred term (*Tables 3.15.3.1, 3.15.3.2*). Adjudicated death events will be listed in *Listing 7.4*.

6.5.1.5 Paradoxical Bronchospasms

All paradoxical bronchospasm events will be captured on the AE CRF page. Paradoxical bronchospasms will be summarized by treatment during the randomized-treatment period *(Table 3.2.3).*

6.5.2 Clinical Laboratory Measurements

Lab parameters collected include the following:

Table 7 Lab Parameters

Hematology	
Hemoglobin	Mean corpuscular hemoglobin
Hematocrit	Mean corpuscular hemoglobin concentration
White blood cell count with differential	Mean corpuscular volume
Red blood cell count	Eosinophils
Platelet count	
Clinical Blood Chemistry	
Liver Enzyme and Other Liver Function Tests	**Other Clinical Blood Chemistry**
Alanine aminotransferase (ALT)	Albumin
Aspartate aminotransferase (AST)	Blood Urea Nitrogen (BUN)
Alkaline phosphatase	Calcium[a]
Bilirubin, total	Chloride[a]
Gamma-glutamyl transferase	Cholesterol
	Bicarbonate
	Creatinine[a]
	Glucose[a]
	Magnesium
	Potassium[a]
	Phosphate
	Protein, total
	Sodium[a]
	Triglycerides
Urinalysis	
Macroscopic examination including specific gravity, pH, protein, glucose, ketones, blood, and urobilinogen.	
Other Tests:	
Pregnancy test (women of childbearing potential only): serum hCG at Visit 1 (Screening) and Visit 13 (Week 52 or Final Study Visit) or Treatment Discontinuation/Study Withdrawal Visit Creatinine clearance will be estimated by the CKD-EPI formula [Levey, 2009].	

Abbreviations: CKD-EPI=Chronic Kidney Disease Epidemiology Collaboration; hCG=human chorionic gonadotropin

[a] Parameters included in the Basic Metabolic Panel.

A Clinically Significant Laboratory Abnormality as identified by the Investigator after the start of study treatment will be recorded as an Adverse Event and tabulated as an AE in the AE analysis. Abnormalities occurring prior to the start of treatment will be noted in medical history and presented in a data listing. Per protocol, the criteria for a "clinically significant" laboratory abnormality are:

- A laboratory abnormality that leads to a dose-limiting toxicity (e.g., an abnormality that results in study drug dose reduction, suspension or discontinuation)
- A laboratory abnormality that results in any therapeutic intervention (i.e., concomitant medication or therapy)
- Other laboratory abnormality judged by the Investigator to be of any particular clinical concern (e.g., significant fall in hemoglobin not requiring transfusion)

All laboratory data will be stored in the database with the units in which they were originally reported. Laboratory data not reported in International System of Units (SI units; Système International d'Unités) will be converted to SI units before data analysis.

Individual clinical laboratory variables for hematology and clinical chemistry and kidney function, including creatinine clearance, will be provided in listings *(Listing 8.1* for hematology*, Listing 8.2* for blood chemistry and kidney function, *Listing 8.3* for urinalysis, and *Listing 4.6* for pregnancy test results at screening and after the start of treatment). Data will be listed in SI units where available. Comments for laboratory testing will be listed *(Listing 8.4)*. For listings, laboratory values will be flagged as Low or High based on the reference ranges provided by the central laboratory, Covance *(Appendix 4)*.

The baseline measurement for a laboratory parameter will be the last available measurement prior to the start of dosing.

If there are multiple laboratory values for the same parameter at pre-dose of a visit, the last value will be chosen for analysis.

Summary statistics (n, mean, median, standard deviation, minimum, and maximum) for the baseline assessment and for the pre-dose value and change from baseline at each post-baseline visit and end of treatment for scheduled lab assessments of continuous laboratory variables including serum potassium and glucose will be tabulated. "End of Treatment" is defined as the last non-missing assessment during the treatment period. Data from unscheduled visits will not be used for the by-visit summaries but both scheduled and unscheduled-visit are candidates for clinically significant values, for the end-of-treatment summary, and for shift tables. Data from both scheduled and unscheduled visits will be listed. The summaries will be provided by treatment *(Tables 3.16.1* through *3.16.4,* for hematology, blood chemistry, kidney function, and urinalysis, respectively). Shift tables will be produced using the categories defined by the Common Terminology Criteria for Adverse Events (CTCAE) Version 4.03 grades for the Safety Population *(Tables 3.16.5 to 3.16.8)* for hematology, chemistry, kidney function and urinalysis,

respectively) for hematology, chemistry, kidney function and urinalysis, respectively). For these shift tables, for each treatment, the subject's pre-dose grade will be cross-tabulated by the subject's maximum post-baseline grade during the treatment; also, the subject's maximum post-baseline grade during treatment will be tabulated for all baseline grades combined. Percentages of subjects in each maximum post-baseline grade for a treatment will be calculated for each pre-dose grade for the treatment and also for all baseline grades combined. Laboratory abnormal values on-treatment will be flagged as High or Low values based on laboratory reference ranges provided by Covance (found in Appendix 4) as per Pearl, Inc. These flags along with the reference ranges will be provided in the laboratory data listings.

Potentially Clinically Significant Laboratory Values Above/Below a Clinically Relevant Threshold on-treatment, based on CTCAE 4.03 and other criteria, will be identified based on the following thresholds:

Table 8 Potentially Clinically Significant (PCS) Laboratory Parameter Criteria

Parameter	Post-Baseline Criteria
Hematology	
Hemoglobin	<8.0 g/dL (<80 g/L)
	Increase of >40 g/L to a value above the ULN (upper limit of normal)
White Blood Cell Count	<2000/µL
	>35,000/µL
Platelet Count	<50,000/µL
	>999,000/µL
Chemistry	
eGFR-EPI	<30 mL/min/1.73 m^2
AST	>3 x ULN
ALT	>3 x ULN
Alkaline Phosphatase	>5 x ULN
Total Bilirubin	>2 x ULN
Blood Glucose* (random values)	<2.2 mmol/L (<39.6 mg/dL)
	>13.9 mmol/L (>250 mg/dL) if baseline is below 10.0 mmol/L (180 mg/dL), >16.7 mmol/L (>300 mg/dL) if baseline is greater than 10.0 mmol/L (180 mg/dL).
Serum Potassium	<3.0 mmol/L
	>6.0 mmol/L

*CTCAE 4.03 criteria are based on fasting glucose values. However, subjects were not required to fast prior to obtaining blood glucose values.

Potentially clinically significant laboratory values will be tabulated for the Safety Population (*Table 3.16.9*).

Since a reduction in potassium and an increase in blood glucose are known class effects of beta-agonists, all potassium or glucose assessments for subjects who experienced newly occurring or worsening potentially clinically significant values after start of the study treatment will be provided in separate listings *(Tables 3.16.10* to *3.16.11).* For all laboratory parameters other than glucose and potassium noted in Table 7, all laboratory data for the parameter identified as potentially clinically significant for a subject will be listed *(Table 3.16.12*- Safety Population).

6.5.3 Vital Signs

Changes from Baseline in on-treatment supine or seated systolic blood pressure, supine or seated diastolic blood pressure, and heart rate will be evaluated, where baseline is defined as the average of all available pre-dose measurements taken prior to the start of dosing at the Randomization Visit (Visit 3). If there are no Visit 3 pre-dose values, the baseline will be defined as the average of pre-bronchodilator values at Visit 2. No Hypothesis testing will be performed.

A **Clinically Significant Abnormality** in vital signs identified by the Investigator will be recorded as an Adverse Event if it occurs after the start of treatment. These adverse events will be included in the AE summaries; abnormalities prior to the start of treatment will be noted in medical history and listed.

Potentially clinically significant changes in systolic and diastolic blood pressure will be defined based on the following criteria provided by Pearl, Inc.:

Table 9 Potentially Clinically Significant Criteria for Systolic and Diastolic Blood Pressure Parameters

Parameter (mmHg)	Post-Baseline Criteria
Systolic Blood Pressure, increase	≥180 and increase from baseline ≥20
Systolic Blood Pressure, decrease	≤90 and decrease from baseline ≥20
Diastolic Blood Pressure, increase	≥105 and increase from baseline ≥15
Diastolic Blood Pressure, decrease	≤50 and decrease from baseline ≥15

PCS changes in heart rate will be assessed as follows:

Table 10 Potentially Clinically Significant Criteria for Heart Rate Parameters

Parameter	Post-Baseline Criteria
Tachycardia Event	≥110 bpm and increase ≥15% from baseline
Bradycardia Event	≤50 bpm and decrease ≥15% from baseline

Vital sign measurements (Heart rate, systolic blood pressure, diastolic blood pressure and body temperature, weight, height) during the study will be displayed in a vital signs listing *(Listing 9.1)*.

A summary of baseline weight, height, and BMI will be presented by treatment *(Tables 1.4.1 to 1.4.4* for the mITT, RVU, and Safety Populations, and all subjects not randomized, respectively).

Summary statistics (n, mean, median, standard deviation and range) of the absolute value and change from baseline for systolic blood pressure, diastolic blood pressure, and heart rate will be tabulated by treatment, visit, and time point. These summaries (*Table 3.17.1.1, Table 3.17.1.2*) will be prepared for baseline and each scheduled post-baseline nominal time point at each scheduled post-baseline visit and end of treatment. End of Treatment will be summarized for each scheduled post-baseline time point (pre-dose 1 hr, and post-dose 30 minutes and 4 hours). "End of Treatment" for each of these assessment points is defined as the last non-missing on-treatment assessment available for the time point. Data from unscheduled visits will not be used for the by-visit summaries but both scheduled and unscheduled-visit data are candidates for clinically significant values and for the end-of-treatment summary. Data from both scheduled and unscheduled visits will be listed. Time windows will be derived for each post-baseline visit using the time intervals for the study time windows detailed in Table 11. No hypothesis tests will be performed.

Table 11 Analysis Study Time Windows for Vital Signs Assessments

Calculated Study Time Window	Time Interval for the Study Time Window
Pre-dose 60 min.	≥45 min. prior to dose
Post-dose 30 min.	>0 to <75 min. post-dose

Note that minutes are rounded to the nearest whole number before applying time windows.

If there are multiple vital sign values for the same parameter at pre-dose assessments after Visit 3 or within the same post-dose study time window at a visit, the last value will be chosen for analysis.

The percentage of subjects with potentially clinically significant values for vital signs at any time post-dose at a visit will be summarized by treatment based on the criteria in Table 9 and Table 10 (*Table 3.17.2.1*- Safety Population).

All vital sign assessments for subjects with potentially clinically significant values will be listed *(Tables 3.17.3.1, 3.17.3.2, 3.17.4.1, Table 3.17.4.2)*.

6.5.4 12-Lead Electrocardiogram Measurements

Changes from baseline in Heart Rate, PR Interval, QRS Axis, QRS Interval, QT Interval and QTcF (Fridericia Corrected QT) interval will be calculated where baseline is defined as the average of the pre-dose measurements taken prior to the start of treatment at the randomization visit (Visit 3). If there are no Visit 3 pre-dose values, the baseline will be defined as the value obtained at Visit 1. The QTcF is defined as $[QT/(RR^{1/3})]$. Heart rate (bpm) is estimated as 60,000/RR, where RR is in units of ms. These assessments will be tabulated for each treatment and assessment time.

A **Clinically Significant Abnormality** for a 12-Lead ECG measurement identified by the Investigator as a clinically significant abnormality will be recorded as an Adverse Event if it occurred after the start of study treatment. These adverse events will be included in the AE summaries.

All 12-Lead ECG measurements for the Safety Population will be listed *(Listing 9.2).* Summary statistics (mean, median, standard deviation and range) for raw values and change from baseline values in Heart Rate, PR Interval, QRS Axis, QRS Interval, QT Interval and QTcF interval will be calculated. These assessments will be tabulated for each treatment and each scheduled nominal time point at each visit and at end of treatment (*Table 3.18.1*). End of Treatment is defined as the last non-missing on-treatment assessment available. Data from unscheduled visits will not be used for the by-visit summaries but both scheduled and unscheduled-visit data are candidates for clinically significant values and for the end-of-treatment summary. Data from both scheduled and unscheduled visits will be listed. Mean pre-dose change from baseline for heart rate and QTcF will be plotted across post-baseline visits by treatment *(Figures 3.18.1 to*

3.18.5). ECG data from subjects with pacemakers will not be included in analyses, but will be listed.

If there are multiple ECG values for the same parameter at pre-dose of a visit date (other than for Visit 3), the last value will be chosen for analysis.

Other than for the change from baseline analyses mentioned above, all available data post-baseline including data from unscheduled visits will be used for ECG parameter analyses.

Table 12 Criteria for PCS ECG Values

Parameter	Post-Baseline Criteria
QTcF Prolongation	(1) ≥ 500 msec if < 500 msec at study baseline and ≥ 15 msec change from study baseline (2) ≥ 530 msec if ≥ 500 msec at study baseline and ≥ 15 msec change from study baseline (3) ≥ 500 msec and ≥ 15 msec change from study baseline (4) Change of ≥60 msec from study baseline regardless of initial value

Potentially clinically significant ECG parameter values will be identified based on criteria listed in Table 12. The number and percentage of subjects who had such values observed any time post-dose will be tabulated for each treatment (*Table 3.18.2*) and listed (*Table 3.18.3)* for QTcF prolongation). No hypothesis tests will be performed.

6.5.5 Healthcare Resource Utilization

Data on healthcare resource utilization will be collected at all visits post-baseline and summarized by treatment group.

The following variables will be calculated unadjusted (per subject) over the entire Treatment Period and tabulated by actual treatment received for those subjects for whom they or one or more of their family members missed work:

- The number of days missed work due to COPD.
- The number of days that caregivers of subjects missed from work as a result of the subject's COPD.

The following variables will be tabulated by actual treatment received and relationship to COPD (COPD-related, not COPD-related, and combined). The mean and the mean per person-year will be calculated across all subjects in a treatment.

- The percentage of subjects with telephone calls to health-care providers:
  - Calls to any health-care provider (physician or other)
  - Calls to physician

  - Calls to other health-care provider

- The mean number of telephone calls to health-care providers:
  - Calls to any health-care provider (physician or other)
  - Calls to physician
  - Calls to other health-care provider

- The percentage of subjects with visits to health-care providers:
  - Visits to any health-care provider (GP, specialist, or other)
  - Visits to GP
  - Visits to specialist
  - Visits to other health-care provider

- The mean number of visits to health-care providers:
  - Visits to any health-care provider (GP, specialist, or other)
  - Visits to GP
  - Visits to specialist
  - Visits to other health-care provider

- Ambulance Transport
  - The percentage of subjects who required ambulance transport
  - The mean number of times ambulance transport was required

- ER Visits
  - The percentage of subjects with ER visits
  - The mean number of visits to ERs

- Hospitalizations
  - The percentage of subjects hospitalized
  - The mean number of subject hospitalizations
  - The mean number of days in the hospital

- Hospitalizations with some time spent in the ICU or CCU
  - The percentage of subjects hospitalized with some time spent in the ICU or CCU
  - The mean number of subject hospitalizations with some time spent in the ICU or CCU
  - The mean number of days in the hospital with some time spent in the ICU or CCU

- Hospitalizations with No time spent in the ICU or CCU
  - The percentage of subjects hospitalized with No time spent in the ICU or CCU
  - The mean number of subject hospitalizations with No time spent in the ICU or CCU

  - The mean number of days in the hospital with No time spent in the ICU or CCU

- ICU
  - The percentage of subjects in the ICU
  - The mean number of days in ICUs

- CCU
  - The percentage of subjects in the CCU
  - The mean number of days in CCUs

Analyses will be performed using the mITT Population.

Descriptive statistics (n, mean, standard deviation, median, minimum and maximum) will be provided by actual treatment received for the number of days missed from work per year, the number of days that family members of subjects missed from work per year overall during the study *(Table 3.20.1 and Listing 9.4)*.

Also, descriptive statistics will be provided by actual treatment received and relationship to COPD (related, not-related, and total) overall during the entire Treatment Period for the following variables: the number of telephone calls to health-care providers, the number of visits to health-care providers, the number of ER visits, the number of number of times ambulance transport was required, the number of subject hospitalizations, the number of days in the hospital, the number of days in the ICU, and the number of days in the CCU (*Table 3.20.2 and Listings 9.4 and 9.5*).

6.5.6 Physical Examination

Any physical examination abnormality reported after the start of treatment for a subject is to be reported as an adverse event. Thus, these will be included in listings of adverse events and summarized in adverse event summaries. Abnormalities seen at the Screening physical examinations will be recorded as Medical History and listed.

7. CHANGES FROM METHODS PLANNED IN THE PROTOCOL

The changes described in this section are discrepancies from protocol version 4.0 (Amendment 3).

- The Type I error control figure in the protocol for the ex-US approach (Protocol, Figure 9-4) incorrectly lists “RS-Total Score” as a secondary analysis. It should be “EXACT-Total Score”. The correction has been made to the figure in this SAP (Figure 8).

Any further changes to methods planned in this SAP will be documented in a revision to this statistical plan prior to database lock, or identified in the clinical study report.

8. STATISTICAL SOFTWARE

Data processing, statistical screening, descriptive reporting and analysis of the efficacy and safety data will be performed using SAS (Version 9.3 or higher). Graphs may also be produced using R (R Development Core Team, 2003).

9. REFERENCES

Carroll KJ, Fleming TR. Statistical Evaluation and Analysis of Regional Interactions: The PLATO Trial Case Study. *Statistics in Biopharmaceutical Research 2013;* 5:2, 91-101. DOI: 10.1080/19466315.2013.783878
COPD Assessment Test (CAT) (http://www.catestonline.org/)

Cro, S (2017) Relevant Accessible Sensitivity Analysis for Clinical Trials with Missing Data. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: 10.17037/PUBS.03817571 (http://researchonline.lshtm.ac.uk/3817571/1/2017_EPH_PhD_CRO_S.pdf).

EuroQol Group. EQ-5D. Accessed at http://www.euroqol.org/about-eq-5d/how-to-use-eq-5d.html (November, 2014).

Farrar JT, Dworkin RH, Max MB. Use of the cumulative proportion of responders analysis graph to present pain data over a range of cut-off points: Making clinical trial data more understandable. J Pain Symptom Manage 2006; 31:369-377

Hochberg Y, A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75, 4, pp. 800-2.

Japanese Respiratory Society (JRS). Guidelines for the Diagnosis and Treatment of COPD, 4th edition, 2013.

Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604-612.

Little RJA, Rubin DB. Statistical analysis with missing data, (2nd edn.). John Wiley and Sons, Inc.: New York, 2002.

Molenberghs G, Kenward MG. Missing data in clinical studies. John Wiley and Sons, Inc.: New York, 2007.

National Center for Health Statistics. 1996. NHANES III Reference Manuals and Reports. Data Dissemination Branch, Hyattsville, MD. CD-ROM No. 6-0178 (1096).

National Research Council. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The prevention and treatment of missing data in clinical trials. The National Academies Press: Washington, DC, 2010.

R Development Core Team. R: A language and environment for statistical computing, 2003. R Foundation for Statistical Computing, Vienna, Austria.

Ratitch B, O'Kelley M, and Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. Pharmaceutical Statistics, 2013; 12: 337-347.

Rubin DB. M. Multiple imputation for nonresponse in surveys. John Wiley and Sons, Inc.: New York, 1987.

SAS Institute Inc. 2008. SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Seaman SR, White IR, Leacy FP. Comment on "Analysis of Longitudinal Trials With Protocol Deviations: A Frameword for Relevant, Accessible Assumptions, and Inference via Multiple Imputation," by Carpenter, Roger, and Kenward. Journal of Biopharmaceutical Statistics, 2014; 24(6):1358-1362.

Van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health 15, 2012 (708-715).

van Reenan M, Janssen B. EuroQol Group EQ-5D-5L User Guide: Basic information on how to use the EQ-5D-5L instrument. https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf (April 2015)

White IR, Royston P, and Wood AM. Multiple imputation using changed equations: issues and guidance for practice. Statistics in Medicine, 2011; 30: 377-399.