AstraZeneca

Statistical Analysis Plan

Study Code	PT010005
NCT #	NCT02465567
Date	30 April 2019

A Randomized, Double-Blind, Multi-Center, Parallel Group Study to Assess the Efficacy and Safety of PT010 Relative to PT003 and PT009 on COPD Exacerbations over a 52-Week Treatment Period in Subjects With Moderate to Very Severe COPD

STATISTICAL ANALYSIS PLAN FOR STUDY PT010005

Protocol Numbers: PT010005

Investigational Drug and Drug Number: BGF MDI; PT010
GFF MDI; PT003
BFF MDI; PT009

Indication: COPD

Dosage Form/Dose:
- BGF MDI 320/14.4/9.6 µg ex-actuator BID
- BGF MDI 160/14.4/9.6 µg ex-actuator BID
- GFF MDI 14.4/9.6 µg ex-actuator BID
- BFF MDI 320/9.6 µg ex-actuator BID

PT010005 Protocol Title: A Randomized, Double-Blind, Multi-Center, Parallel Group Study to Assess the Efficacy and Safety of PT010 Relative to PT003 and PT009 on COPD Exacerbations over a 52-Week Treatment Period in Subjects With Moderate to Very Severe COPD

Date of Issue: 30 April 2019

Version: Version 2.0

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 (YYYY-MM-DD)**	**Reason for the Change / Revision**
Version 2.0	2019-04-30	"Other" region has been removed from description of individual regions (US & Canada, Asia, Western Europe, Eastern Europe, Mexico & Central America & South America, Australia & New Zealand & South Africa) since existing region designations are all-inclusive.
Version 2.0	2019-04-30	Additional details on the COPD exacerbation analyses were added to outline what would be done if convergence is not obtained.
Version 2.0	2019-04-30	Updated the attributable estimand methodology to use a bootstrap approach. It has been shown that Rubin's rules are conservative for the attributable estimand, and a bootstrap method is generally more powerful while controlling the type I error rate using the maximum likelihood approach of von Hippel. This approach does not require the taking of posterior draws from a posterior distribution. Note that for tipping point analyses for the Efficacy Estimand and the Treatment Policy Estimand, Rubin's rules will continue to be used.
Version 2.0	2019-04-30	Updated wording on rules for medication classification in order to improve clarity.
Version 2.0	2019-04-30	Clarified that the start day of a COPD exacerbation will not be excluded from the time-at-risk of a COPD exacerbation. This was done to avoid a logical inconsistency.
Version 2.0	2019-04-30	To add "baseline post-bronchodilator percent predicted FEV_1" as a covariate for the responder analysis for TDI. This makes the TDI responder analysis consistent with the main analysis of TDI focal score as a continuous endpoint. It also makes the TDI responder analysis more consistent with the SGRQ responder analysis. The mention of this covariate was inadvertently omitted in the version 1.0 of the SAP.
Version 2.0	2019-04-30	For FEV_1 AUC $_{0\text{-}4}$, version 1.0 of the SAP text had inadvertently stated that the analysis would be ANCOVA. Now it is clarified that this is a linear mixed model repeated measures ANCOVA analysis (which is the analysis type also being used for predose trough FEV_1). This brings the SAP text in line with how the SAP mocks for the main FEV_1 $AUC_{0\text{-}4}$ analysis have already been.
Version 2.0	2019-04-30	For rate of decline in pre-dose trough FEV_1, baseline value will not be included in the model as time point zero, but will be included as a covariate in the model. The reason is so that only post-baseline changes are reflected in the rate. Likewise, for rate of decline in FEV_1 $AUC_{0\text{-}4}$, baseline value will be included as a covariate in the model.

Version 2.0	2019-04-30	Criteria for upper limit for Glucose were changed for potentially clinically significant laboratory values. Criteria for eGFR-EPI were changed for potentially clinically significant laboratory values. The reason was to incorporate revised sponsor guidelines, which had been changed at least in part due to revisions in the CTCAE guidelines from CTCAE version 4.03 to 5.0.
Version 2.0	2019-04-30	If the COPD exacerbation resulted in a hospitalization or death and there is no known start date of prescribed treatment with a systemic corticosteroid or systemic antibiotic for the exacerbation, then the start date of the exacerbation will be the date of such hospitalization or death (whichever is first). For any moderate or severe COPD exacerbation, if the end date of the exacerbation is unknown, then the end date of the exacerbation will be assumed to be nine days after the start date of the exacerbation, for a default duration of 10 days. A CRF-recorded COPD exacerbation resulting in a hospitalization or death will be considered as a countable moderate to severe (and also severe) COPD exacerbation regardless of whether systemic corticosteroids or systemic antibiotics were used to treat the exacerbation. The reason is to avoid the exclusion of meaningful COPD exacerbation records from the analysis.
Version 2.0	2019-04-30	Updated text to reflect that the [redacted] iSAP were completed prior to the time of the SAP amendment.
Version 2.0	2019-04-30	Changed order of secondary objectives to be consistent with the protocol. To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on symptoms of COPD is the first secondary objective. Version 1.0 of the SAP had inadvertently listed this objective in a different position.
Version 2.0	2019-04-30	Updated 24-hour Holter monitoring sub-study endpoints to reflect previous Holter study endpoints from similar studies.
Version 2.0	2019-04-30	Added imputation rule for baseline blood eosinophil count. The reason is to avoid the exclusion of meaningful data from the analysis (on the basis of missing covariates).
Version 2.0	2019-04-30	Specified how to handle subjects who were re-screened for the baseline post-bronchodilator percent predicted FEV_1. The reason is to avoid ambiguity.
Version 2.0	2019-04-30	Added baseline percent reversibility to bronchodilator to Baselines and Baseline Covariates for Analysis section. The reason is so that a subject's values can still be used if Ventolin reversibility is missing but Atrovent reversibility is not missing.
Version 2.0	2019-04-30	Updated analyses to include the log baseline blood eosinophil count as a covariate instead of baseline blood eosinophil count. The reason for the logarithmic transformation is that eosinophils counts are skewed right and transformation reduces the leverage of a few high values. Also provided a rule by which the log baseline eosinophil count can be derived if the baseline eosinophil count is zero.
Version 2.0	2019-04-30	Removed analysis of the following other endpoints out of SAP to reduce the volume of output for the CSR and to facilitate review: • Compliance for EQ-5D • PFT sub-study endpoints: FVC, PEFR, and FEF_{25-75} • Individual components of the TDI: functional impairment, magnitude of task, and magnitude of effort over 24 weeks, over 52 weeks, and at each post-randomization visit • Individual domains of the SGRQ: Symptoms, Activity, and Impacts over 24 weeks, over 52 weeks, and at each post-randomization visit • EXACT symptom domain scores for breathlessness, cough and sputum, and chest symptoms over 24 weeks, over 52 weeks and over each 4-week interval of the 52-week treatment period • Rate and time to first COPD exacerbation treated with systemic steroids • Rate and time to first COPD exacerbation treated with antibiotics Likewise, the eDiary-data listings and the listing of individual-question SGRQ data have been removed.

Version 2.0	2019-04-30	Added subgroup analysis for safety section.
Version 2.0	2019-04-30	Created Appendix 9: operational definitions for exacerbations. Moved some text out of the main body of the SAP and into that appendix.
Version 2.0	2019-04-30	Seasonal analyses were removed from SAP text to reduce volume of output for the CSR and to facilitate review.
Version 2.0	2019-04-30	Added more start and end date imputation rules to the current duration of COPD exacerbation rules section. If the end date of a COPD exacerbation is precisely known but the start date is unknown or only partially known, then the start date will be imputed assuming a 10-day duration, within the constraints of what is possible based on the partial date. If the start and end dates are both partially missing where the month and year are known, the earliest 10-day window that is consistent with the partial dates will be imputed. If only the year is known, the exacerbation will not be counted. The reason is to avoid excluding meaningful COPD exacerbation records from the analysis.
Version 2.0	2019-04-30	GOLD ABCD categories were removed from subgroup analysis. This is because the categories were outdated.
Version 2.0	2019-04-30	Rate of severe COPD exacerbations is now a secondary endpoint instead of an "other" efficacy endpoint and will be analyzed for 3 estimands (Efficacy, Attributable, and Treatment Policy). The reason is that this study's main objectives pertain to COPD exacerbations, and this is an important measure of efficacy.
Version 2.0	2019-04-30	Some sections of the Details Appendix (Appendix 6) were re-ordered or re-grouped in order to increase clarity. Also, this Appendix now refers to bootstrapping and the maximum likelihood approach of von Hippel instead of Rubin's rules for the Attributable Estimand.
Version 2.0	2019-04-30	In the SAS Code Appendix (Appendix 3), "Other" was dropped as a region because no subjects were enrolled in countries other than those already assigned to explicitly named regions.
Version 2.0	2019-04-30	To ensure that endpoints for which a decrease is favorable (including SGRQ, Rescue Ventolin HFA Use, and EXACT) will incorporate such directionality in the cumulative responder analysis, as applicable.
Version 2.0	2019-04-30	Stipulated that for treatment discontinuations, COPD exacerbations that start one day after the last treatment date are considered to be on-treatment COPD exacerbations. The reason is to avoid the exclusion of meaningful and relevant COPD exacerbations.
Version 2.0	2019-04-30	Worded the definitions of prior, concomitant, and post-treatment medications more clearly.
Version 2.0	2019-04-30	Stipulated that calculation of FEV_1 AUC_{0-4} for the efficacy, treatment policy, and per-protocol estimands will require at least one non-missing post-dose value. The reason is to provide more detail about the calculation.
Version 2.0	2019-04-30	Revised the magnitude of the non-inferiority margins of the following endpoints: rescue Ventolin HFA, TDI, FEV_1 AUC_{0-4}, SGRQ total score, SGRQ (MCID) response, and EXACT total score. The changes are as follows: rescue Ventolin HFA (changed from 0.5 to 0.75 puffs/day), TDI (changed from 0.5 to 0.75) (which is still less than the MCID of 1), FEV_1 AUC_{0-4} (changed from 50 to 75 mL), SGRQ total score (2 to 3) (which is still less than the MCID of 4), SGRQ (MCID) response (changed from 5 to 10 percentage points), and EXACT total score (changed from 1 to 1.5). These revisions are made for consistency with other similar recent studies and for removing unnecessary conservatism while still providing margins that are less than the MCID for endpoints for which an MCID has been defined.
Version 2.0	2019-04-30	Clarified the negativity of the non-inferiority margins for continuous endpoints, so that the margin is negative for endpoints for which an increase is beneficial (including TDI focal score, pre-dose trough FEV_1, and FEV_1 AUC_{0-4}), whereby the margin is applied to the lower confidence bound of the 95% confidence interval for the treatment difference (BGF 160 minus BFF).
Version 2.0	2019-04-30	Clarified the positivity of the non-inferiority margins for continuous (or binary) endpoints, so that the margin is positive for endpoints for which a decrease is beneficial (including rescue use, SGRQ total score, SGRQ (MCID) response, and EXACT total score), whereby the margin is applied to the upper confidence bound of the 95% confidence interval for the treatment difference (BGF 160 minus BFF).

Version 2.0	2019-04-30	Clarified that an adverse event that begins on the day after the date of premature treatment discontinuation is still considered to be on-treatment (i.e. treatment-emergent).
Version 2.0	2019-04-30	Clarified that a death is considered to be on-treatment (i.e. treatment-emergent) if any AE leading to death is on-treatment (i.e. treatment-emergent).
Version 2.0	2019-04-30	Revised the definition of "Time to Onset of Action Assessed Using FEV_1 on Day 1" to be in terms of a mean instead of a median.
Version 2.0	2019-04-30	For cumulative responders analyses, made the distinction that the subject is considered to be a non-responder if premature discontinuation from treatment occurred prior to the time point or time- interval-end (for the time point or time interval to which the analysis pertains). This is to ensure that patients who discontinued after the analysis time interval are not classified as non-responders.
Version 2.0	2019-04-30	Provided more detail about the quantities to be graphed in the Q-Q plots.
Version 2.0	2019-04-30	Added the definition of the category of "mild" COPD. The reason is for comprehensiveness.
Version 2.0	2019-04-30	Updated the section about Adverse Events of Special Interest to take into account further medical review.
Version 2.0	2019-04-30	Added secondary objective: to assess the effect of BGF MDI relative to GFF MDI and BFF MDI on COPD exacerbations to cover additional COPD-exacerbation-related endpoints not covered by the primary objective.
Version 2.0	2019-04-30	Provided more clarity for the censoring rules for time to death, time to treatment failure, and time to first moderate or severe COPD exacerbation.
Version 2.0	2019-04-30	To add more detail to the mock tables and listings (in Appendix 6 of the SAP: Details Appendix) that will aid in a blinded categorization of discontinuation reasons as being MAR/MCAR, or MNAR and a blinded categorization of discontinuation reasons as Attributable or Not Attributable (to treatment).
	2019-04-30	Added cumulative responder analyses for pre-dose trough FEV_1 and FEV_1 AUC_{0-4} for over 24 weeks.
	2019-04-30	Added the following safety analyses to SAP text and mocks: • "Overall Summary of AEs" and "Incidence of AEs by SOC and Preferred Term" and "Incidence of Confirmed Pneumonia" and "Incidence of Confirmed MACE" that cover the following time intervals: 0 to <=24 Weeks and >24 Weeks. • "Incidence of Adverse Events Leading to Death by SOC and preferred term". Do one table for on-treatment (treatment-emergent) AEs and another table for post-treatment AEs.
	2019-04-30	Added text to clarify calculation of the information fraction, and how it will be used to determine the alpha level for the final analyses.
	2019-04-30	If the baseline COPD exacerbation history (categorical covariate) is missing, then it will be set equal to 1 or >=2 based on the information entered into the IWRS system for the COPD exacerbation history stratum.
	2019-04-30	Added non-inferiority tipping point analysis table and figure for rate of moderate or severe COPD exacerbations for the per-protocol estimand for BGF 160 versus BFF.

TABLE OF CONTENTS

List of Tables

List of Figures

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

AE	Adverse event
ALT	Alanine aminotransferase
ANCOVA	Analysis of covariance
AR(1)	Autoregressive order 1
ARH(1)	Heterogeneous Autoregressive order 1
AST	Aspartate aminotransferase
AUC_{0-4}	Area under the curve from 0 to 4 hours
BID	Bis in die
BDI	Baseline Dyspnea Index
BFF MDI	Budesonide and Formoterol Fumarate Metered Dose Inhaler
BGF MDI	Budesonide, Glycopyrronium, and Formoterol Fumarate Metered Dose Inhaler
BID	Bis in die, twice daily
bpm	Beats per minute
[redacted]	[redacted]
CAT	Chronic Obstructive Pulmonary Disease Assessment Test
CEC	Clinical Endpoint Committee
CCU	Coronary Care Unit
CCV	Cardio- and cerebrovascular
CD	Compact disc
CI	Confidence interval
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
cm	Centimeter
COPD	Chronic obstructive pulmonary disease
CSR	Clinical study report
CTCAE	Common Terminology Criteria for Adverse Events
DBP	Diastolic blood pressure

DMC	Data Monitoring Committee
E-RS	EXACT Respiratory Symptoms
ECG	Electrocardiogram
eCRF	Electronic Case Report Form
eDiary	Electronic Diary
e.g.	Exempli gratia; for example
eGFR	Estimated glomerular filtration rate
EQ-5D	EuroQol 5 Dimensions Questionnaire
EQ-5D-5L	EuroQol 5 Dimensions Questionnaire 5-level
ER	Emergency Room
ex-actuator	Dose delivered from the actuator (i.e., mouthpiece) of the MDI
ex-US	Outside the United States
EXACT	Exacerbations of Chronic Pulmonary Disease Tool – Patient Reported Outcomes
FEV_1	Forced expiratory volume in the first second
FEF_{25-75}	Forced expiratory flow between 25% and 75% of FVC
FVC	Forced vital capacity
GFF MDI	Glycopyrronium and Formoterol Fumarate Metered Dose Inhaler
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GP	General practitioner
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
HR	Heart rate

ICF	Informed Consent Form
ICS	Inhaled corticosteroid
ICU	Intensive Care Unit
i.e.	Id est; that is
IQR	Interquartile range
ISAP	Interim statistical analysis plan
ITT	Intent-to-Treat
IWRS	Interactive Web Response System
λ	Rate of COPD exacerbations
JRS	Japanese Respiratory Society
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
MCID	Minimal clinically important difference
MDI	Metered dose inhaler
MedDRA	Medical Dictionary for Regulatory Activities
µg	Microgram
MI	Multiple Imputation
mITT	Modified Intent-to-Treat
mL	Milliliter
mm	Millimeter
mmHg	Millimeter of mercury
MNAR	Missing Not at Random
msec (ms)	Millisecond

NHANES	National Health and Nutrition Examination Survey
OCS	Oral corticosteroid
OTC	Over-the-counter
PMM	Pattern Mixture Model
PCS	Potentially clinically significant
PEFR	Peak expiratory flow rate
PFT	Pulmonary function test
PP	Per-Protocol
PT	Preferred Term
PT003	Glycopyrronium and Formoterol Fumarate Inhalation Aerosol
PT009	Budesonide and Formoterol Fumarate Inhalation Aerosol
PT010	Budesonide, Glycopyrronium, and Formoterol Fumarate Inhalation Aerosol
PVC	Premature ventricular contraction
QoL	Quality of Life
QTcF	QT corrected using Fridericia's formula
RVU	Rescue Ventolin User
RM	Repeated measures
ROM	Read-only memory
SABA	Short-acting β_2-agonist
SAE	Serious adverse event
SAMA	Short-acting muscarinic antagonist
SAP	Statistical Analysis Plan
SBP	Systolic blood pressure
SD	Standard deviation
SGRQ	St. George's Respiratory Questionnaire
SMQ	Standard MedDRA Query
SOC	System Organ Class

sPDP	Statistical protocol deviation plan
SVT	Sustained ventricular tachycardia
TC	Telephone call
TDI	Transition Dyspnea Index
TEAE	Treatment-emergent adverse event
TOEPH	heterogeneous Toeplitz
ULN	Upper limit of normal
US	United States
VAS	Visual analog scale
WHO-DD	World Health Organization Drug Dictionary

Trademark Information

Trademarks Not Owned By Pearl

Atrovent

SAS

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 PT010005. Pearl Therapeutics is a member of the AstraZeneca group of companies. Sponsor means either Pearl or AstraZeneca. The SAP should be read in conjunction with the study protocol and the electronic Case Report Form (eCRFs) for the study. This version of the SAP has been developed using the PT010005-03 Amended Protocol (Version 6.0 dated March 27, 2018) and the CRF Revision 2.0 dated July 30, 2015. This is the main SAP, describing the statistical analyses that will be carried out at the end of the study. Separate statistical analysis plans were provided for the [redacted] SAP and Interim SAP (ISAP). They are to be used for reference only since their analyses are completed.

2. STUDY OBJECTIVES AND ENDPOINTS

2.1 Study Objectives

The overall objective of Study PT010005 is to assess the efficacy and safety of BID (Bis in die, twice daily) treatment with BGF MDI 320/14.4/9.6 µg (budesonide, glycopyrronium, and formoterol fumarate metered dose inhaler) and BGF MDI 160/14.4/9.6 µg relative to GFF MDI 14.4/9.6 µg (glycopyrronium and formoterol fumarate metered dose inhaler) and BFF MDI 320/9.6 µg (budesonide and formoterol fumarate metered dose inhaler) over 52 weeks in subjects with moderate to very severe chronic obstructive pulmonary disease (COPD).

2.1.1 Primary Objective

- To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on the rate of moderate or severe COPD exacerbations.

2.1.2 Secondary Objectives

- To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on symptoms of COPD
- To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on quality of life (QoL)
- To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on all-cause mortality
- To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on COPD exacerbations

2.1.3 Safety Objectives

- To assess the safety of BGF MDI relative to GFF MDI and BFF MDI.

2.1.4 Sub-Study Objectives

2.1.4.1 4-Hour Pulmonary Function Test (PFT) Sub-study Objective

- To assess the effect of BGF MDI relative to GFF MDI and BFF MDI on lung function

2.1.4.2 24-Hour Holter Monitoring Sub-study Objective

- To evaluate the cardiovascular safety of BGF MDI, GFF MDI, and BFF MDI as evaluated by 24-hour Holter monitoring

2.1.4.3 Health Care Resource Utilization Objective

- To assess overall and COPD-specific Healthcare Resource Utilization (HCRU) of BGF MDI, GFF MDI, and BFF MDI

2.2 Study Endpoints

2.2.1 Efficacy Endpoints

2.2.1.1 Primary Efficacy Endpoint

- Rate of moderate or severe COPD exacerbations

2.2.1.2 Secondary Efficacy Endpoints

- Time to first moderate or severe COPD exacerbation
- Change from baseline in average daily rescue Ventolin Hydrofluoroalkane (HFA) use over 24 weeks
- Transition Dyspnea Index (TDI) focal score over 24 weeks (outside the United States [US], ex-US)
- Change from baseline in the Exacerbations of Chronic Pulmonary Disease Tool – (EXACT) total score over 52 weeks (ex-US)
- Change from baseline in St. George's Respiratory Questionnaire (SGRQ) total score over 24 weeks (ex-US)
- Percentage of subjects achieving a minimal clinically important difference (MCID) of 4 units or more in SGRQ total score at Week 24 (US)
- Time to death (all cause)
- Rate of severe COPD exacerbations

2.2.1.3 Other Efficacy Endpoints

Wherever stated, analyses of an endpoint at each post-randomization visit will be performed only at time points where the endpoint will be assessed per the schedule of assessments.

- Rate of COPD exacerbations of any severity
- Time to first COPD exacerbation of any severity
- Time to first severe COPD exacerbation
- Change from baseline in average daily rescue Ventolin HFA use over 52 weeks
- Time to death (respiratory)
- Time to treatment failure (treatment discontinuation for any cause, moderate or severe exacerbation, or death)
- Change from baseline in: the EXACT total score, the EXACT Respiratory Symptoms (E-RS) total score over 24 weeks, over 52 weeks and over each 4-week interval of the 52-week treatment period
- Percentage of days with "no rescue Ventolin® HFA use"
- Other TDI endpoints:
    - TDI focal score over 24 weeks (US), over 52 weeks, and at each post-randomization visit
    - Percentage of subjects achieving an MCID threshold of 1 unit or more on average over 24 weeks in TDI focal score
    - Percentage of subjects achieving an MCID threshold of 1 unit or more on average over 52 weeks in TDI focal score
- Other SGRQ endpoints:
    - Change from baseline in SGRQ total score over 52 weeks, and at each post-randomization visit
    - Percent of subjects achieving an MCID of 4 units or more in SGRQ total score over 52 weeks, at Week 52, and over 24 weeks.
- EuroQol 5 Dimensions Questionnaire 5-level (EQ 5D-5L) variables including the EuroQol 5 Dimensions Questionnaire (EQ-5D) index score, the EQ-5D Visual Analog Scale (VAS), and five-dimension single item 5-level responses at each post-randomization visit.

2.2.2 Safety Endpoints

The safety endpoints include:

- Adverse events (AEs)
- 12-lead electrocardiograms (ECGs)
- Clinical laboratory testing
- Vital signs measurements

2.2.3 Sub-Study Endpoints

2.2.3.1 4-Hour Pulmonary Function Test Sub-Study Endpoints

- Change from baseline in morning pre-dose trough FEV_1 (forced expiratory volume in the first second) at Week 24 (US) and over 24 weeks (ex-US) for the comparison of BGF MDI to GFF MDI
- FEV_1 area under the curve from 0 to 4 hours (AUC_{0-4}) at Week 24 (US) and over 24 weeks (ex-US) for the comparison of BGF MDI to BFF MDI

Other PFT Sub-Study Endpoints:

- Change from baseline in morning pre-dose trough FEV_1 over 24 weeks, over Weeks 12 to 24, over 52 weeks and at each post-randomization visit
- FEV_1 AUC_{0-4} over 24 weeks, over Weeks 12 to 24, over 52 weeks and at each post-randomization visit where measured
- Peak change from baseline in FEV_1 over 24 weeks, over Weeks 12 to 24, over 52 weeks and at each post-randomization visit where measured
- Rate of decline in pre-dose FEV_1 over 52 weeks
- Rate of decline in FEV_1 AUC_{0-4} over 52 weeks
- Time to onset of action on Day 1

2.2.3.2 24-Hour Holter Monitoring Sub-Study Endpoints Assessed at Visit 8 (Week 16)

Primary 24-Hour Holter Monitoring Sub-Study Endpoints Endpoint

- Change from baseline in mean heart rate averaged over 24 hours

Secondary 24-Hour Holter Monitoring Sub-Study Endpoints

- Change from baseline in mean nighttime (22:00 to 06:00) and daytime (06:00 to 22:00) heart rate
- Change from baseline in the maximum 24-hour heart rate
- Change from baseline in the minimum 24-hour heart rate
- Change from baseline in the frequency of isolated ventricular ectopic events (premature ventricular contractions [PVCs])
- Change from baseline in the frequency of ventricular couplets (defined as two PVCs preceded or followed by regular beats)
- Change from baseline in the frequency of ventricular runs (defined as three or more PVCs preceded or followed by regular beats)
- Change from baseline in the frequency of supraventricular ectopic beats
- Change from baseline in the frequency of isolated supraventricular ectopic events
- Change from baseline in the frequency of supraventricular couplets

- Change from baseline in the frequency of supraventricular runs
- Incidence of withdrawal criteria being met during 24-hour Holter monitoring
- Incidence of atrial fibrillation with rapid ventricular response (>100 beats per minute [bpm])

Other 24-hour Holter Monitoring Sub-study Endpoints

- Proportion of subjects with maximum heart rate > 180, > 160 - ≤ 180, > 140 - ≤ 160, > 120 - ≤ 140, > 100 - ≤ 120, and 100 bpm or less
- Proportion of subjects with minimum heart rate > 60, > 50 - 60, > 40 - 50, and ≤ 40 bpm
- Proportion of subjects in each category of change from baseline in the number of PVCs per hour (no change, increase of > 0 - < 60, ≥ 60 - < 120, and ≥ 120, and decrease of > 0 - < 60, ≥ 60 - < 120, and ≥ 120)

2.2.4 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 healthcare 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
  - Call 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 subject 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 subject spent No time in the ICU or the CCU
- The percentage of subjects hospitalized with No time 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 ICU
- The percentage of subjects in the ICU
- The mean number of days in 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

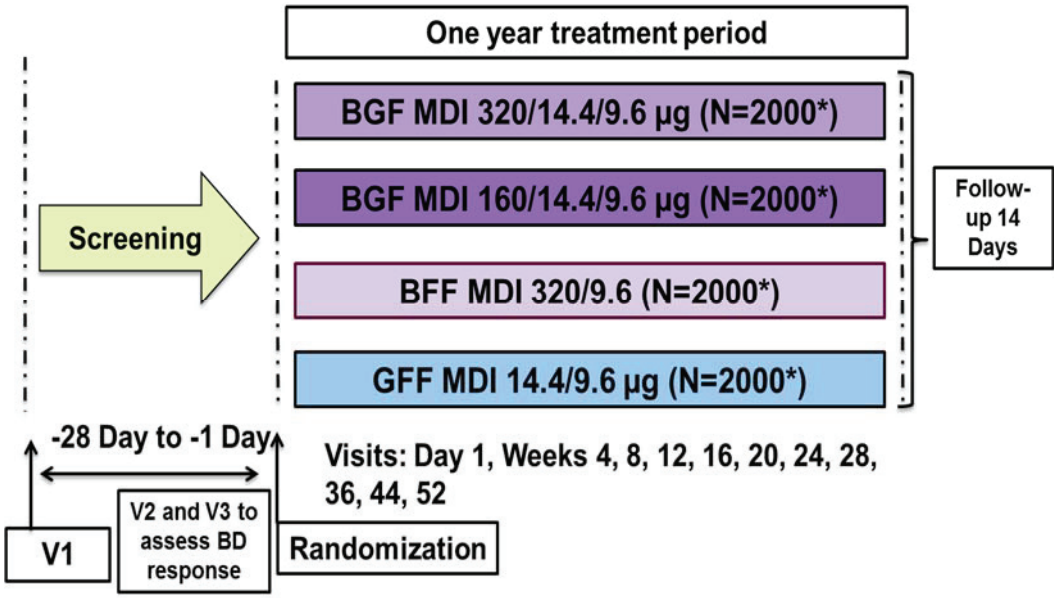
3. STUDY DESIGN AND ANALYTICAL CONSIDERATIONS

3.1 Study Design

This is a randomized, double-blind, multi-center, parallel group, group sequential study to assess the efficacy and safety of BGF MDI 320/14.4/9.6 µg (micrograms) and BGF MDI 160/14.4/9.6 µg relative to GFF MDI 14.4/9.6 µg and BFF MDI 320/9.6 µg over a 52-week treatment period in approximately 8,400 subjects with moderate to very severe COPD with an increased risk of experiencing a COPD exacerbation and that remain symptomatic (CAT ≥ 10, where CAT denotes chronic obstructive pulmonary disease assessment test) on two or more inhaled maintenance treatments. There was a blinded sample size reassessment where the underlying rate of moderate or severe COPD exacerbations and negative binomial dispersion parameter were predicted and the sample size increased from 8,000 to 8,400. The sample size reassessment was conducted prior to the interim efficacy analysis. The one efficacy interim analysis was conducted when 50% of the information for the study was accrued, and over 500 subjects were randomized in China. If overwhelming efficacy for both doses had been observed at the interim efficacy analysis or for the high dose with a clear dose response, the study would have been terminated at that time.

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

Figure 1 Study Design



Abbreviations: BFF=Budesonide and Formoterol Fumarate; BGF=Budesonide, Glycopyrronium, and Formoterol Fumarate; GFF=Glycopyrronium and Formoterol Fumarate; MDI=metered dose inhaler; *=sample size may be increased; BD=bronchodilator.

3.1.1 Overall Study Design and Plan

Subjects who successfully complete the Screening Period will then be randomized in a 1:1:1:1 ratio to BGF MDI 320/14.4/9.6 µg BID, BGF MDI 160/14.4/9.6 µg BID, BFF MDI 320/9.6 µg BID, or GFF MDI 14.4/9.6 µg BID, respectively.

Approximately 2,100 subjects were to be randomized to each treatment arm. Randomization was stratified by exacerbation history (1 or ≥ 2 moderate or severe exacerbations), post-bronchodilator FEV_1 (25% to < 50% or 50% to 65% predicted), blood eosinophil count (< 150 cells per mm^3 or ≥ 150 cells per mm^3), and country. With protocol Version 1.0 (dated 18 May 2015) up to a 1:2 ratio for the blood eosinophil strata was targeted with twice as many randomized subjects in the ≥ 150 cells per mm^3 category. With protocol Amendment 5 (Version 6.0 dated 27 Mar 2018) for the last portion of enrollment, the eosinophil target was changed. Subjects with a blood eosinophil count of < 150 cells per mm^3 or ≥ 150 cells per mm^3 at Visit 1 became eligible to participate in the main study; however, the PFT sub-study remained closed to subjects with an eosinophil count of < 150 cells per mm^3 at Visit 1.

Subjects who discontinue study treatment prior to Week 52 (Visit 14) will be encouraged to remain in the study to complete all remaining study visits during the 52 week treatment period. All subjects who agree to continue study participation beyond treatment discontinuation will complete a Treatment Discontinuation/Withdrawal Visit prior to transitioning back to regularly scheduled study visits. Subjects participating in the Holter sub-study who choose to discontinue

from treatment will complete only regularly scheduled visits for the main study and will not complete any remaining sub-study Holter assessments. Subjects participating in the PFT sub-study who agree to post-treatment-discontinuation follow-up will continue serial PFTs as scheduled. Treatment discontinuation subjects will return to appropriate maintenance COPD medications, per the investigators 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.

Sub-Studies:

This study includes 2 sub-studies, a PFT sub-study, and a 24-hour Holter Monitoring sub-study. A subset of sites will be identified and designated for participation in the PFT- and 24-Hour Holter monitoring sub-studies. Individual subjects will be allowed to participate in both sub-studies.

A subset of approximately 3,060 subjects (765 subjects per treatment arm) at designated study sites will be invited to participate in the PFT sub-study. In these subjects, serial PFTs will be conducted over 4 hours at selected visits throughout the 52-week Treatment Period for selected visits and collection times.

A subset of approximately 800 subjects (200 subjects from each treatment group) at designated study sites will be invited to participate in the 24-Hour Holter Monitoring sub-study at selected visits.

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

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

All prescription and OTC medications taken by the subject during 30 days before Visit 1 (Screening) will be recorded on the prior/concomitant medications eCRF. All concomitant medications taken during the study will be recorded on the Concomitant Medications eCRF page with indication, total daily dose, route of administration, and start/stop dates of drug administration.

3.2 Hypothesis Testing

For the primary comparisons, the null hypothesis for each pairwise comparison will be that the rate of moderate or severe COPD exacerbations of subjects taking BGF MDI (2 doses: 320/14.4/9.6 µg and 160/14.4/9.6 µg) is equal to that of subjects taking either GFF MDI or BFF MDI; the alternative hypothesis is then that the rate of moderate or severe COPD exacerbations of subjects taking BGF MDI (2 doses: 320/14.4/9.6 µg and 160/14.4/9.6 µg) is less than that of subjects taking either GFF MDI or BFF MDI. P-values will thus be reported as one-sided. The exception is that in addition non-inferiority of BGF MDI 160/14.4/9.6 µg to BFF MDI will be testing prior to testing for superiority.

The primary null (H_0) and alternative (H_1) hypotheses with λ representing the mean number of moderate or severe COPD exacerbations are:

- H_0: $\lambda_{BGF320/14.4/9.6} = \lambda_{GFF}$
  H_1: $\lambda_{BGF320/14.4/9.6} < \lambda_{GFF}$

- H_0: $\lambda_{BGF320/14.4/9.6} = \lambda_{BFF}$
  H_1: $\lambda_{BGF320/14.4/9.6} < \lambda_{BFF}$

- H_0: $\lambda_{BGF160/14.4/9.6} = \lambda_{GFF}$
  H_1: $\lambda_{BGF160/14.4/9.6} < \lambda_{GFF}$

- H_0: $\lambda_{BGF160/14.4/9.6} / \lambda_{BFF} > 1.1$
  H_1: $\lambda_{BGF160/14.4/9.6} / \lambda_{BFF} \leq 1.1$

Secondary and other efficacy analyses will involve the above hypotheses applied to secondary efficacy endpoints. The directionality – being "< "or ">" -- of H_1 will depend on the endpoint.

3.3 Sample Size

[redacted]

Under these assumptions and based on 8,400 subjects randomized in a 1:1:1:1 ratio for BGF MDI 320/14.4/9.6 µg, BGF MDI 160/14.4/9.6 µg, BFF MDI, and GFF MDI, respectively, [redacted]

[redacted]

3.4 Interim Analysis

The sample size reassessment was conducted prior to the interim efficacy analysis. [redacted] If overwhelming efficacy for both doses had been observed at the interim efficacy analysis or for the high dose with a clear dose response, the study would have been terminated at that time. This interim efficacy analysis was conducted under the auspices of a Data Monitoring Committee (DMC) and is described in a separate interim SAP (ISAP).

[redacted]

Because the study is continuing until all subjects complete, the one-sided alpha criteria for the final analysis will be adjusted based on the percent of information available at the prior interim analyses such that the overall Type I error is controlled at one-sided 0.025.

The total information at the end of the study will be calculated based on the following formula from Page 6 in Mutze et al (Mutze et al., 2018) using parameter estimates from the primary analysis using the negative binomial model after unblinding. The overall negative binomial rate of moderate or severe COPD exacerbations will be used for μ, and the dispersion parameter will be used for Φ. The time at risk for each subject, *i,* at the final analysis will be used for t_i and these will be summed over all (n) subjects in the mITT population across all treatment groups.

$$I = \sum_1^n \frac{t_i \mu}{1 + \phi t_i \mu}$$

The information accrued at the time of the interim analysis will also be calculated using the above formula using the interim analysis overall negative binomial exacerbation rate estimate for μ, and the interim analysis dispersion parameter estimate for Φ. The time at risk at the interim will be used for *t* and these will be summed over all subjects in the interim mITT population. This information accrued at the interim will then be divided by the total information at the end of the study to arrive at the information fraction.

The information fraction will be used to calculate the level of statistical significance for the final analysis such that the overall level of significance α is maintained at a one-sided level of 0.025. Critical values on the p-value scale will be obtained using SAS 9.4 (or higher) PROC SEQDESIGN for a two-stage, one-sided superiority test with α = 0.025, with information levels set to the observed information fraction at interim, $\hat{\omega} = (\hat{I}_{interim}/\hat{I}_{final})$, using the formula above to calculate observed information at the interim and final, and 1. The power family error spending function will be used:

$$E(t;\rho) = \begin{cases} 1 & if\ t \geq 1 \\ t^{\rho} & if\ 0 < t < 1 \\ 0 & otherwise \end{cases}$$

Set $\rho = \ln(0.2)/\ln(\hat{\omega})$, so that 20% of the total error is spent at the observed interim information fraction: $E(\hat{\omega}, \rho) = 0.2$.

The final-analysis critical level of significance will be reported in the CSR *(Table 2.2.4)* and discussed in the statistical methods appendix to the CSR.

In addition to the above interim efficacy analysis, a DMC will also review safety data approximately every 6 months. A separate document, the interim statistical analysis plan (ISAP), separate from this main SAP, describes the interim analyses which will be provided to the DMC at both the safety reviews and the efficacy interim and describes the stopping rules at the efficacy interim.

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.

5. ANALYSIS POPULATIONS

5.1 Population Definitions

5.1.1 Intent-to-Treat (ITT) Population

The **ITT 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 randomized treatment group. Efficacy 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 **mITT Population** is a subset of the ITT Population, defined as all subjects with post-randomization data obtained prior to discontinuation from the study treatment. Any data obtained after completion of or discontinuation from the study treatment will be excluded from the mITT analysis but will still be included in the ITT analysis. Subjects will be analyzed according to randomized treatment group. 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 except for the non-inferiority analyses. Note: The knowledge that a subject did not have a COPD exacerbation constitutes an efficacy assessment.

5.1.3 Rescue Ventolin User Population

Regional differences in rescue Ventolin HFA usage are expected with subjects in some regions using virtually no rescue medication at study entry. Therefore, 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 Per-Protocol (PP) Population

The **PP Population** is a subset of the mITT Population defined as all subjects with post-randomization data obtained prior to any major protocol deviations. Data obtained after any major protocol deviation or discontinuation from treatment will be excluded. Since receiving the wrong treatment is a major protocol deviation, data after the deviation from such subjects will be excluded from the PP Population and will be analyzed as randomized (which for this population is identical to analysis by the actual treatment received). Any evaluability criteria with a potential impact on efficacy results will be identified during Blinded Data Review Meeting (BDRM) prior to database lock. Major protocol deviations, therefore, can result in exclusion of all data from a particular subject from the PP population or require exclusion of data from a specific time point and/or subsequent time points for an endpoint.

Criteria for exclusion from the PP population will be established at a BDRM prior to database lock. Reasons for exclusion from the PP Population will include, but are not limited to, the following:

1. An incorrect diagnosis of COPD.
2. Subjects who do not have an established clinical history of COPD and severity.
3. Subjects who do not have a demonstrated history of COPD exacerbations.
4. For those subjects who require rescue Ventolin HFA less than 6 hours before study visit, all data post-Ventolin HFA administration will be considered missing for that day.
5. For spirometry endpoints, the subject will not be eligible for the endpoint-specific, visit-specific PP population if the subject did not take study medication in the evening prior to the visit day.

6. For subjects who take any protocol-prohibited medication that would affect spirometry assessments on the date of an assessment, spirometry measurements taken at that assessment will be excluded from the per-protocol analysis.

The PP Population will be the primary population for all non-inferiority efficacy comparisons.

5.1.4.1 Record-Level Exclusion of PFT data from the PP Population

Participation in the PFT sub-study is defined as being enrolled in the PFT sub-study and having at least one post-baseline PFT assessment. Data obtained after treatment discontinuation will be excluded from the PP Population for PFT endpoints.

Also, PFT records are excluded from the PP Population if the following condition is true:

- Subjects who cannot meet protocol-specified baseline stability criteria. FEV_1 baseline stability is defined as the average of the 60-minute and 30-minute pre-dose FEV_1 assessments at Visit 4 being within ±20% or 200 mL of the mean of the pre-bronchodilator FEV_1 assessments obtained at the 2 preceding visits (average of pre-dose FEV_1 assessments obtained at Visit 2 and Visit 3).

Subjects who fail to meet any of the three restriction criteria prior to spirometry will be handled as follows. The 3 restrictions are (1) subject was not to smoke for at least 4 hours prior to study visit and throughout the duration of each study visit, (2) subject was not to use xanthine-containing products (i.e., coffee, tea, cola and chocolate) for at least 6 hours prior to study visit and for the duration of each study visit, and (3) subject was not to have COPD bronchodilator medications for at least 6 hours prior to study visit. Spirometry data for these subjects at the affected visits will be removed from the PP Population for PFT endpoints. Such restrictions will be applied only if data pertaining to the meeting of the criteria (including the timing) were collected.

Subjects who did not take the evening dose of study medication on the day prior to a visit will be excluded from the visit-specific PP population for PFT endpoints.

5.1.5 Safety Population

The **Safety Population** is similar to the mITT Population (all subjects who are randomized to treatment and receive any amount of the study treatment). However, 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 had no post-dose safety assessments. (Note that 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.

5.1.6 Holter Monitoring Population

The **Holter Monitoring Population** is defined as all subjects in the Safety Population who had at least 18 hours of acceptable quality Holter monitoring data at both Visit 3 (Holter Monitor Baseline) and Visit 8 (Week 16). Exclusions from this population may be identified by Pearl Therapeutics prior to database lock and unblinding.

5.1.7 PFT Sub-Study Membership

A subject is considered to be "in" the PFT sub-study if the subject was randomized and any of the following three criteria are met:

1. The CRF has "yes" as the response to the question "Is the subject participating in the PFT sub-study" and has "yes" as the response to the question "Did subject meet FEV_1 baseline stability".

2. The spirometry data have "yes" as the response to the question "Confirm subject is participating in the PFT sub-study".

3. The subject has at least one post-baseline spirometry assessment (after the first dose of study medication).

When the PFT sub-study subset (of subjects) is used for analysis, it will always be used in the context of some study population, such as the mITT Population, the ITT Population, or the PP Population.

5.2 Populations for Primary and Sensitivity Analyses

Demographics will be summarized for the mITT, PP, RVU, and Safety Populations as well as for subjects participating in the 4-hr PFT and Holter sub-studies. Extent of exposure will be summarized for the Safety Population and for the subjects participating in the PFT and Holter sub-studies. The Safety Population will be used to summarize safety. The mITT Population will be used for the primary efficacy analyses, with the ITT and PP populations being considered supportive with a couple exceptions. The ITT Population will be used for a secondary analysis for the US approach, and the PP Population will be considered to be the primary population for any non-inferiority analyses. Analyses of Holter monitoring parameters will be performed using the Holter Monitoring Population. Selected analyses will be performed for specified sub-populations. See the Subgroup Analyses (Section 6.4.8).

This study had a planned efficacy interim analysis which was completed prior to the SAP amendment and which did not result in a decision to terminate the study for unequivocal efficacy. The plan for that interim analysis was described in a separate SAP, Had the interim-analysis decision been to terminate the study for overwhelming efficacy, the analysis for the rate of moderate or severe exacerbations performed at the interim would have served as the primary analysis. Were that to have occurred, the main analyses for selected secondary efficacy variables analyzed at the interim would still have included data through study closeout and final database

lock. Because the study did not terminate at the interim analysis, the main analysis for both the primary and secondary endpoints will include data through the study closeout and final database lock.

6. STATISTICAL ANALYSIS

All data collected on the CRFs and contributing to the analysis will be provided in listings, except for data collected only for confirmation of study entry criteria and for study management purposes and except for individual-question data for the SGRQ. 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. eDiary data listings will not be included (due to the size of the database).

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 of the 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 definitely related. Imputed values will not be listed in data listings.

Data Imputation 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 the end-of-treatment summary, 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.

On-treatment COPD exacerbations

An exacerbation will be considered "on-treatment" if its start date is after the first study treatment and before or on the date of last study treatment. For premature treatment discontinuations, this definition is extended to include exacerbations starting one day after the last treatment date.

6.2 Subject Disposition and Analysis Populations

A disposition table for all subjects randomized will be provided (*Table 1.1.1*). This 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, who completed the study, and who completed the study and completed follow-up phone call. The number and percentage of randomized subjects included in the mITT, ITT, PP, RVU, Safety, Holter Monitoring Populations, and PFT-sub-study subject membership in the mITT, ITT, and PP Populations will also be tabulated (*Table 1.1.1*). Informed consent is listed in *Listing 9.7*.

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 and cross-classification of reversibility to Ventolin HFA and disease severity, using IWRS data and Clinical data for the stratification-factor levels, will be tabulated in *Tables 1.1.5.1* to *1.1.6.2*. If the clinical-data-based exacerbation history is missing, the IWRS exacerbation history stratum category (1 or >=2) will be used. 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 will also be listed (*Listing 1.3)*. A list of the discrepancies between IWRS-based and clinical-data-based stratification factors will be provided (*Listing 1.6*).

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*). Reasons for premature discontinuation from study treatment will be summarized for all subjects randomized (*Table 1.2.1*).

The number and percentage of subjects in the mITT and PP Populations who discontinued study treatment for each reason for discontinuation of treatment will be tabulated by randomized treatment (*Table 1.2.2* and *1.2.4*) and in the Safety Population, by treatment actually received (*Table 1.2.1*).

The number and percentage of subjects in the ITT Population who withdrew from the study and each reason for subjects' withdrawal from the study will be tabulated (*Table 1.2.6*).

Membership in various analysis populations will be listed by subject *(Listing 4.10)*. The reason for exclusion from the PP Population will be tabulated by study treatment for all mITT subjects (*Table 1.3*). The reason for exclusion of a subject from the ITT, mITT, PP, and Safety Populations or exclusion of partial data (at some but not all time points) for a subject will be listed for all randomized subjects (*Table 1.1.4*). A listing of subjects who did not comply with restrictions on smoking, use of rescue medication, and xanthine-containing products (protocol deviations requiring removal of data from the PP Population analysis) just prior to spirometry will be provided in *Listing 6.1.1*.

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, Physical Characteristics, CAT

Subject demographics and baseline characteristics will be summarized for the mITT, PP, RVU, Safety Population, and for Non-Randomized subjects (*Tables 1.4.1* through *1.4.5*, 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. If the Safety Population has the same treatment assignment as the mITT, then it will be identical as well and hence not tabulated. Demographics will also be summarized for the subjects in the PFT sub-study (the ITT Population for PFT sub-study subjects) and the Holter Monitoring Population (*Tables 1.4.6* and *1.4.7*). The mITT population does not need to be tabulated because it is the same as the ITT population for demographics; this is true overall and for the PFT Sub-study.

The COPD Assessment Test (CAT) is a short self-administered questionnaire designed to assess the condition of subjects and overall impact of COPD. It has been proven that the CAT has good repeatability and discriminative properties which suggest that it is sensitive to treatment effects at a group level. CAT will be done at Screening Visit 1 and Visit 4 prior to the first dose of study medication. It will be utilized to describe the burden and symptomatic impact of COPD in the study populations for those subjects who are enrolled into the clinical trial.

Demographic and baseline characteristic variables summarized will include the following:

- Age (summarized as continuous variable and by the following categories: < 65 years and >= 65 years)
- Duration of COPD

- Gender
- Race
- Ethnicity (each category from the CRF will be employed)
- Weight
- Height
- BMI (kg/m^2)
- Smoking history/status (non-smoker, current smoker, and former smoker)
  - Former smoker is defined as those who have stopped smoking for at least 6 weeks prior to first Screening visit.
- Used inhaled corticosteroids at Screening (all populations except for Non-Randomized subjects)
- Baseline eosinophil count (as continuous variable and by the following categories: <150 cells per mm^3 vs. ≥150 cells per mm^3)
- Baseline exacerbation history (1 vs. ≥2)
- 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
- COPD Assessment Test (CAT) total score and total score category at screening and baseline (<10, ≥10, <15, ≥15, <20, ≥20)
  - The baseline CAT score is the last non-missing value prior to the first dose of study medication.

Screening and baseline CAT data will also 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 for all subjects for the mITT and Safety Populations by treatment and listed (*Tables 1.5.1* and, *1.5.2*, and *Listing 4.1*). Severity of COPD at Screening based on post-bronchodilator data at Visit 2 (if not missing) or at Visit 3 if the data are missing at Visit 2 will also be included in these summaries. History of moderate or severe COPD exacerbations in the prior year will be summarized and listed (*Tables 1.9.3* and *1.9.4* for the Safety and mITT Populations, respectively, and *Listing 4.3*). Exacerbations occurring during the screening period will be included in this summary and listing.

Descriptive statistics will be provided for screening period pre-bronchodilator and post-bronchodilator and baseline spirometry parameters (*Tables 1.6.1, 1.6.3, 1.6.4,* and *1.6.5* for the mITT, PP, Safety Population, and PFT sub-study subjects, respectively, and *Listing 2.2*).

Characterization of Reversibility:

Reversibility to Ventolin HFA (Short-acting β_2-agonist, SABA) will be evaluated at Visit 2. Reversibility to Atrovent HFA (short-acting anticholinergic) will be evaluated at Visit 3. Spirometry is conducted approximately 60 minutes and 30 minutes prior to bronchodilator administration and approximately 30 minutes post-bronchodilator administration at these visits.

"Reversibility to bronchodilator" is the reversibility to Ventolin HFA if reversibility to Ventolin HFA is not missing; otherwise, reversibility to bronchodilator is reversibility to Atrovent HFA.

"Percent reversibility to bronchodilator" is the percent reversibility to Ventolin HFA if percent reversibility to Ventolin HFA is not missing; otherwise, percent reversibility to bronchodilator is percent reversibility to Atrovent HFA. This is the percent reversibility that will be used as a covariate in the applicable efficacy analyses.

Reversibility (%) to bronchodilator is defined as 100 x (the change from pre-bronchodilator HFA to post for FEV_1)/pre-bronchodilator HFA FEV_1. Reversible (Yes/No) is defined as improvement in FEV_1 post-bronchodilator HFA administration compared to pre-bronchodilator HFA of $\geq$ 12% and $\geq$ 200mL.

Reversibility to Ventolin HFA at Screening Visit 2 and reversibility to Atrovent HFA at Screening Visit 3 will be summarized for the mITT, PFT sub-study, and PP Populations (*Tables 1.7.1, 1.7.2,* and *1.7.3* for Ventolin HFA reversibility, and *1.8.1*, *1.8.2,* and*1.8.3* for Atrovent HFA reversibility). The 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. These data will also be listed (*Listing 2.2*). Atrovent HFA and Ventolin HFA dispensing will also be listed (*Listing 5.2*).

Additionally, the percentage of subjects meeting each of the following response criteria will be included in the tabulations for both Ventolin HFA and Atrovent bronchodilators:

- $\geq$ 12% improvement post-bronchodilator in FEV_1 from pre-bronchodilator
- $\geq$ 150 mL improvement post-bronchodilator in FEV_1 from pre-bronchodilator
- $\geq$ 200 mL improvement post-bronchodilator in FEV_1 from pre-bronchodilator

6.3.3 Medical and Surgical History at Screening, Cardiovascular Medical History, and 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 1.9.1* 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 1.9.2* 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 over-the-counter (OTC) medications taken by the subject during 30 days before Screening will be recorded on the Concomitant Medications case report form (CRF) page.

Coding: Verbatim medication/treatment terms will be coded by [redacted] and will be assigned a preferred term (PT) and an ATC (anatomic therapeutic class) term using the latest version of the World Health Organization Drug Dictionary (WHO-DD) available (version: 3Q 2016 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. All prior medication taken by the subject within 30 days of Screening for the study and all concomitant therapy taken by the subject while on study will be recorded in the eCRF.

Prior medication/treatment is any medication that was used at any time prior to the date of first 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 that was used at any time between the date of first study treatment and the date that is one day before the date of discontinuation from or completion of study medication.

Any medication that was used at any time on or after the day of treatment completion or treatment discontinuation is considered a **Post-Treatment medication/treatment**.

If the date of first study medication is missing, then the date of randomization is used.

Otherwise, in the case of missing dates, any period (Prior, Concomitant, and/or Post-Treatment) that is possible based on the available information should be applied.

Reported prior medications for COPD, COPD-exacerbation-related, and non-COPD-related medications will be tabulated *(Tables 1.10.1.1, 1.10.2.1,* and *1.10.3.1*, respectively.

Prior COPD medications will be tabulated (for the Safety population) for subjects having received one, two, and all three, or none of the following treatments, whether in fixed combination products or as separate inhalers, at the time of screening and separately for post-treatment (for any duration): (1) a muscarinic antagonist, (2) a β2 agonist, and (3) an inhaled corticosteroid (*Table 1.10.1.2*). 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.

Concomitant COPD, COPD-exacerbation-related, and non-COPD-related medications/treatments will be summarized by preferred term and actual treatment received for the Safety Population (*Tables 1.11.1, 1.11.2, 1.11.3, 1.11.4.1, 1.11.4.2, 1.11.5.1,* and *1.11.6.1*). Separate tabulations will be made for concomitant medications taken during study treatment and concomitant medications taken during post-treatment period (which can also called the post-treatment-discontinuation follow-up). COPD-related medication summaries will not include the COPD-exacerbation medications.

Prior, concomitant/post-treatment COPD, COPD-exacerbation-related, and non-COPD-related medications will also be displayed in separate listings (*Listings 4.7, 4.8,* and *4.9,* respectively*)*.

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 day 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 1.12.1* for the Safety Population, *Table 1.12.2* for the PFT sub-study Population, and *Table 1.12.3* for the Holter Monitoring Population).

In addition, treatment compliance will be provided in these summaries. 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 in *Listing 9.6*.

6.4 Efficacy Analyses

There are four pairwise comparisons of treatments of interest, namely, BGF 320 vs. BFF, BGF 320 vs. GFF, BGF 160 vs. BFF, and BGF 160 vs. GFF. Unless otherwise stated as for the purpose of testing non-inferiority, the comparisons of interest will be performed for the purpose of testing superiority. All comparisons will be for superiority, with the exception of BGF MDI 160/14.4/9.6 µg MDI to BFF, which will be for non-inferiority first, followed by superiority.

Non-inferiority analyses will chiefly use the PP population, unless specifically stated otherwise. Additionally, the comparisons of GFF to BFF and BGF 320 to BGF 160 will be presented on most analysis tables.

6.4.1 Estimands

The primary estimand of interest is called 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 three additional estimands of interest. One is called the attributable estimand and 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 an unfavorable outcome. Another estimand of interest is called the treatment policy estimand. This estimand is the effect of randomized treatment over the study period regardless of whether randomized treatment is continued. The final estimand of interest is called the per-protocol estimand. This estimand is the effect of treatment on subjects who are compliant with the protocol (i.e. no major protocol deviations), including the use of randomized medication.

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, using a bootstrapped multiple imputation approach (von Hippel, 2017). 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 95th percentile of the reference arms' distribution if the reason is reasonably attributable to tolerability or lack of efficacy. The 95th percentile applies to an endpoint for which a higher value is a worse outcome; however, the 5th percentile applies to an endpoint for which a higher value is a better outcome. More detail about the computation of the attributable estimand will be provided in subsequent sections (especially 6.4.4.1 and 6.4.7.1) and in the Details Appendix to this SAP (Appendix 6).

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 possibly, probably, or definitely 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.

Finally, the last estimand of interest is the per-protocol estimand. Analysis of this estimand will use the PP Population.

6.4.2 Baselines and Baseline Covariates for Analysis

For the diary symptom score parameters and rescue medication usage, baseline will be the mean 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 category is set to 1 or ≥ 2 moderate or severe exacerbations in the last 12 months (from the Visit 1 CRF page) or during the screening period. Cases with 0 historical exacerbations will be categorized as "1" for the purpose of this baseline covariate. If the baseline COPD exacerbation history is missing, then it will be set equal to the COPD exacerbation history stratum (from the IWRS system) that was used for subject randomization.

Baseline percent predicted FEV_1 is the mean of the 30-minute and 60-minute values of percent predicted FEV_1 prior to dosing on Day 1 (Visit 4) if such values are available. However, if there are no such assessments on Day 1, then the mean of the 30-minute and 60-minute pre-bronchodilator assessments at Visit 3 will be employed. If there are no such assessments on Day 1 or Visit 3, then the mean of the 30-minute and 60-minute pre-bronchodilator assessments at Visit 2 will be employed.

The baseline value for any PFT (spirometry) parameter is the mean of the 30 minute and 60 minute values prior to dosing on Day 1 (Visit 4) if such values are available. However, if there are no such assessments on Day 1, then the mean of the 30-minute and 60-minute pre-bronchodilator assessments at Visit 3 will be employed. If there are no such assessments on Day 1 or Visit 3, then the mean of the 30-minute and 60-minute pre-bronchodilator assessments at Visit 2 will be employed.

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

- the subject was taking a medication that contained a component (active ingredient) that is 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). Medications may be but are not necessarily ongoing medications.

Note that WHO denotes World Health Organization.

Baseline blood eosinophil count is the mean of all blood eosinophil count values prior to the first dose of study medication. If the subject is missing pre-IMP eosinophil count data, then the first non-missing post-baseline value will be used as the baseline value. The log baseline eosinophil count (a baseline covariate for analysis) is the natural logarithm of the baseline blood eosinophil count; however, if the baseline blood eosinophil count is zero, then the log baseline eosinophil count will be set to the natural logarithm of 1, which is zero.

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

Baseline post-bronchodilator FEV_1 is the highest value of FEV_1 obtained 30 minutes after dosing with Ventolin at Visit 2 (or later if there are repeated assessments) or, if the Visit-2 value is missing, the highest value of FEV_1 obtained 30 minutes after dosing with Atrovent at Visit 3.

The baseline post-bronchodilator percent predicted FEV_1 is the baseline post-bronchodilator FEV_1 divided by the predicted FEV_1 and multiplied by 100. If a subject is re-screened, the information from the later screening will be employed in the calculation.

Baseline percent reversibility to Ventolin is 100 x (POST-PRE)/PRE, where POST is the highest value of FEV_1 obtained 30 minutes (or later if there are repeated assessments) after dosing with Ventolin at Visit 2 and PRE is the mean of the -30 minute and -60 minute values of FEV_1 prior to dosing with Ventolin at Visit 2.

"Baseline percent reversibility to bronchodilator" is the percent reversibility to Ventolin HFA if percent reversibility to Ventolin HFA is not missing; otherwise, percent reversibility to bronchodilator is percent reversibility to Atrovent HFA. This is the percent reversibility that will be used as a baseline covariate for statistical analyses.

6.4.3 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.

For efficacy analysis or derivation of AUC based on time points, the change from baseline in PFT assessments will be allocated to derived nominal collection time windows using the time intervals specified for each below.

Table 1 Analysis Study Time Window for Spirometry Assessments

Calculated Study Time Window	Time Interval for the Study Time Window
Pre-dose 60 min.	≥ 45 minutes prior to dose
Pre-dose 30 min.	≥ 0 to < 45 minutes prior to dose
Post-dose 5 min.	> 0 to 9 min. post-dose
Post-dose 15 min.	10 to 22 min. post-dose
Post-dose 30 min.	23 to 44 min. post-dose
Post-dose 1 hr.	45 to 89 min. post-dose
Post-dose 2 hrs.	90 to 179 min. post-dose
Post-dose 4 hrs.	3 to < 5 hrs. post-dose

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

If there are multiple spirometry values for the same parameter within the same post-baseline study time window on the same day, the last value will be chosen for analysis.

6.4.4 Primary Efficacy Analyses

Analyses for the primary endpoint are presented in this section along with analyses for any of the secondary or other efficacy endpoints related to the primary endpoint. Calculation of FEV_1 AUC_{0-4} for the efficacy, treatment policy, and per-protocol estimands will require at least one non-missing post-dose value.

6.4.4.1 Rate of Moderate or Severe COPD Exacerbations

A COPD exacerbation will be defined as a change in the subject's usual COPD symptoms that lasts 2 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, and fever without other cause).

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

- Use of systemic corticosteroids and/or antibiotics for at least 3 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 even if the treatment duration was less than 3 days (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) and even if the CRF-recorded exacerbation does not have documented treatment with systemic corticosteroids and/or antibiotics.

- COPD-related death even if the CRF-recorded exacerbation does not have documented treatment with systemic corticosteroids and/or antibiotics

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

Additionally, the investigator may identify certain events (recorded on a linked 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 subsection titled "Duration of COPD Exacerbation."

The rate of moderate or severe COPD exacerbations 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 log baseline blood eosinophil count as continuous covariates, and baseline COPD exacerbation history, region (US & Canada, Asia, Western Europe, Eastern Europe, Mexico & Central America & South America, Australia & New Zealand & South Africa), and ICS use at screening 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. The logarithm of time at risk of experiencing an exacerbation will be used as an offset variable in the model. For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority ratio of 1.1 will be employed.

For the COPD exacerbation analyses, if convergence is not attained the following will be done. First, initial parameter estimates will be provided. If convergence is still not attained, then generalized estimating equations will be used. If, further, convergence is not attained then covariates will be removed from the model one-by-one until convergence is attained, with the order of removal being: region, ICS use at screening, log baseline eosinophil count, post-bronchodilator percent predicted FEV_1, and baseline COPD exacerbation history.

For the efficacy estimand and per-protocol estimand, the time at risk is defined as time of exposure – not during or within 7 days after an exacerbation (of equal or greater severity) – until the last dosing date. However, the start day of a COPD exacerbation will not be excluded from the time at risk. More precisely, this is the amount of time between the date of first dose of study medication and the date of premature discontinuation from study medication (plus one day) or

the date of completion of study medication minus the number of days while the subject was experiencing any exacerbation and minus the seven days subsequent to any exacerbation. An exception is that the start day of the COPD exacerbation is not subtracted from the time at risk. Any days subsequent to the date of discontinuation from or completion of study medication are also not subtracted.

For the attributable estimand, the time at risk is defined as time of exposure or post-exposure not during or immediately subsequent to an actual or imputed exacerbation – until the treatment completion date (for subjects who have completed treatment) or 1 year after the date of first dose (for subjects who have not completed treatment). However, the start day of a COPD exacerbation will not be excluded from the time at risk.

For the treatment policy estimand, time at risk is defined as follow-up time – 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). However, the start day of a COPD exacerbation will not be excluded from the time at risk.

Analyses will be conducted on the mITT Population for the efficacy estimand and attributable estimand, on the ITT Population for the treatment policy estimand, and on the PP Population for the per-protocol estimand. Analyses of the efficacy and treatment policy estimands use only observed data while the attributable estimand requires imputation.

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 the 95th percentile of the reference arms' distribution. The imputed value (using a bootstrap approach with maximum likelihood developed by von Hippel) will be drawn from a negative binomial distribution with mean exacerbation rate (and variance) based on the 95th percentile of the reference arms' distribution, with estimates set to the average of estimates for the two reference treatments from the primary analysis. Other missing data are to be imputed using the observed data model, i.e. assumed to be missing at random (MAR) or missing completely at random (MCAR). The MAR assumption means that missingness is independent of the unobserved outcomes after accounting for the appropriate observed data and covariates in the model. The MCAR assumption means that missingness is independent of the observed and unobserved outcomes after accounting for the appropriate covariates in the model. Further information about the computation of the attributable estimand is given in the Details Appendix to this SAP (Appendix 6). Missingness considered to be attributable will be identified during the BDRM process as outlined at the beginning of Section 6.4.

Sensitivity analyses to explore robustness to missing data for each estimand are described in Section 6.4.4.2.

The number of exacerbations, the percentage of subjects who experience exacerbations, exacerbation rates, the dispersion parameter, and rate ratios comparing treatments will be summarized for moderate or severe exacerbations (*Tables 2.1.1.1, 2.1.1.2, 2.1.1.3,* and *2.1.1.6*

for the efficacy estimand, attributable estimand, treatment policy estimand, and the per-protocol estimand, respectively). The time at risk for moderate or severe COPD exacerbations will also be summarized for the efficacy estimand, treatment policy estimand, and per-protocol estimand (*Tables 2.1.1.1e, 2.1.1.3d,* and *2.1.1.6b)*.

A summary table of the information fraction at the time of the interim analysis will be provided for the efficacy estimand *(Table 2.2.4)*.

An additive-model (or LOESS) plot of the rate of moderate or severe COPD exacerbations versus baseline eosinophil count will also be provided (*Figure 2.1.1.1)*. This analysis will use a generalized additive model (GAM) having a nonparametric regression for the relationship of baseline eosinophil levels to response with a negative binomial model (or will use LOESS). Baseline eosinophil count can be log-transformed for this analysis. The graph will plot the predicted yearly "Rate of Moderate or Severe COPD Exacerbations (events/year)" and its 95% confidence bands on the vertical axis. A semi-logarithmic plot will be made so that the horizontal axis with "Baseline Eosinophil Count (per mm^3)" has logarithmic spacing but has tick labels in the original units (per mm^3). More detail is given in the SAS Code Appendix.

Duration of COPD Exacerbation

For moderate or severe exacerbations, the duration is based on information from the eCRF COPD exacerbation page, whereas for mild exacerbations, the duration is defined by the length of symptoms as recorded in the eDiary.

For moderate or severe COPD exacerbations (using the eCRF COPD exacerbation page), the start date of a moderate or severe COPD exacerbation is the earlier of the medication start date, hospitalization start date (if the exacerbation resulted in hospitalization) and date of death (if the exacerbation resulted in death). The end date of a moderate or severe COPD exacerbation is the later of the medication end date, hospitalization (for exacerbation) end date, and date of death (from the exacerbation).

The duration is the end date minus the start date plus 1 day. 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. If the moderate or severe COPD exacerbation resulted in a hospitalization or death and there is no known start date of prescribed treatment with a systemic corticosteroid or systemic antibiotic for the exacerbation, then the start date of the exacerbation will be the date of such hospitalization or death (whichever is first). For any moderate or severe COPD exacerbation, if the end date of the exacerbation is unknown, then the end date of the exacerbation will be assumed to be nine days after the start date of the exacerbation, for a default duration of 10 days. If the end date of a COPD exacerbation is precisely known but the start date is unknown or only partially known, then the start date will be imputed assuming a 10-day duration, within the constraints of what is possible based on the partial date. If both the start and

end dates are partially missing where the month and year are known, the earliest 10-day window that is consistent with the partial dates will be imputed. If only the year is known, the exacerbation will not be counted.

For mild COPD exacerbations, start date will be defined as the onset of worsened symptoms (lasting at least 2 consecutive days) as recorded by the subject in the electronic diary (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 worsening symptoms (lasting at least 2 consecutive days) or recorded stop date of the earlier event (moderate or severe exacerbation defined by treatment or hospitalization) and start date of the later event (mild, moderate, or severe) will be considered the same event.

In addition, in order to not double-count exacerbations, 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. Note that it is possible for a single “any severity” COPD exacerbation to contain multiple “moderate or severe” COPD exacerbations if there are intervening mild exacerbation symptoms.

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 (i.e. within 7 days of) 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. It will, however, be explained later in this section that moderate and severe COPD exacerbations within 7 days of one another will be coalesced into a single COPD exacerbation event with the severity of “severe”. Note that “i.e.” denotes “id est; that is”.

Assumptions Checks and Removal of Outliers in Sensitivity Analyses

The distribution of residuals and influence statistics will be examined to identify any outliers. In the event that a single, or small number of such outlying values, are found to exist and to be highly influential, the effects may be ameliorated by removal of the outlier as a sensitivity analysis. These analyses will be conducted if warranted to demonstrate the robustness of the results and reported in the Statistical Methods Appendix to the clinical study report (CSR).

6.4.4.2 Sensitivity Analyses for Missing Data: Rate of Moderate or Severe COPD Exacerbations

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, followed by details of the methods.

Table 2 Sensitivity Analyses for Rate of Moderate or Severe COPD Exacerbations

Efficacy Estimand		Attributable Estimand	Treatment Policy Estimand	Per-Protocol Estimand
mITT Population		mITT Population	ITT Population	PP 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 value 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.	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 upper confidence limit for the rate ratio of GFF 160 versus BFF ≥ 1.1.

MNAR = Missing not at random. MNAR means that missingness depends on the unobserved values, and cannot be predicted solely based on the subject's observed data. 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 0.02 exacerbations/year. Imputed values may not be impossible values. Thus the values will be imputed from a truncated distribution.

The primary analysis is for an efficacy estimand and 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

negative binomial regression model with the logarithm of time at risk of experiencing an exacerbation as an offset term and assumes that all missing data is missing at random (MAR) or missing completely at random (MCAR).

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 or MCAR. More detail about the computation of the attributable estimand will be provided in subsequent sections and in the Details Appendix to this SAP (Appendix 6).

Tipping-point analyses will be conducted to examine the impact of varying the rate parameter for missing data in subjects who discontinue BGF MDI. Multiple imputation (MI) techniques will be used to impute the missing data for these patients by varying the exacerbation rate in the BGF MDI arm. The rate in the BGF MDI arm will be increased until the p-value for the comparison of treatment to comparator becomes ≥ 0.05 or until the rate is increased by 1.5 exacerbations/year. 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 Efficacy Estimand:

> Tipping Point #1: this first set of analyses will only impute diminished effects for subjects on BGF MDI whose missing data are determined to be MNAR,
>
> Tipping Point #2: this analysis will impute diminished effects for all missing data in the BGF 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 using the 95th percentile of the reference arms' 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 number of exacerbations that would have been seen is then combined with the observed exacerbations to provide a complete dataset. These data are then analyzed using the same negative binomial model used for the primary analysis. This analysis is repeated multiple times and the results combined using Rubin's formulae [Rubin, 1987]. A total of 10 imputations will be carried out.

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

Tipping Point Analysis of the Per-Protocol Estimand:

For the per-protocol estimand, a non-inferiority tipping point analysis for the rate of moderate or severe COPD exacerbations will be conducted for the comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin for the rate ratio of 1.1 will be employed for the upper bound of the two-sided 95% confidence interval for the rate ratio (for BGF 160 / BFF). A table giving the results for each progressively diminished effect will be produced and a figure of delta (decrement in treatment effect) vs. the upper confidence limit from a two-sided 95% CI will be produced.

6.4.5 Analysis of Secondary Endpoints

Secondary variables include time to first moderate or severe COPD exacerbation, rate of severe COPD exacerbations, TDI focal score over 24 weeks (ex-US), rescue Ventolin HFA usage over 24 weeks, SGRQ total score over 24 weeks (ex-US), percentage of subjects achieving an MCID of 4 units or more in SGRQ total score at Week 24 (US), EXACT total score over 52 weeks (ex-US), and time to death (all cause). Multiplicity will be controlled for the secondary variables as described in Section 6.4.9.

The main analysis for secondary endpoints will use the efficacy estimand with the exception of time to death, which will use the treatment policy estimand. The treatment policy estimand will include all post-treatment discontinuation data and will serve as a sensitivity analysis for the other secondaries. The attributable estimand will also be estimated as a sensitivity analysis for the secondary endpoints with the exception of time to death and the time of onset on Day 1. For the analysis of this estimand, post-treatment discontinuation data are imputed based on the reference arms' distribution if the reason for discontinuation is attributable to lack of efficacy or/tolerability. For rescue Ventolin use, the SGRQ total score, and the EXACT Total Score, the 95th percentile of the reference arms will be used for multiple imputation, while the 5^{th} percentile will be used for the TDI focal score. All other missing data will be imputed using the observed data model. The variances used for the multiple imputation are described in the Details Appendix (Appendix 6). Time to onset on Day 1 is excluded since there is anticipated to be very little missing data at this early time point. Since subjects will often discontinue treatment prior to death, the treatment policy estimand is planned as the main analysis. Also, since providing a bad

outcome of death for attributable missingness is conceptually extreme, the attributable estimand will not be used for this measure.

6.4.5.1 Time to First COPD Exacerbation: Moderate or Severe

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 COPD exacerbation (moderate or severe). Only on-treatment exacerbations will be included for calculating the time to first COPD exacerbation for the efficacy estimand. Exacerbations occurring after the premature discontinuation of treatment will be considered for the treatment policy estimand. For the attributable estimand, a complete dataset will be created analogously to that described in the analysis of Rate of Moderate or Severe Exacerbations in Section 6.4.4.1; 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 1 year.

The time to first moderate or severe COPD exacerbation will be analyzed using a Cox regression model. The model will include treatment, baseline post-bronchodilator percent predicted FEV_1, and log baseline blood eosinophil count as continuous covariates, and baseline COPD exacerbation history ($1, \geq 2$), region (US & Canada, Asia, Western Europe, Eastern Europe, Mexico & Central America & South America, Australia & New Zealand & South Africa), and ICS use at screening (Yes/No) as categorical covariates. Estimated adjusted hazard ratios relative to the comparator will be displayed along with the associated Wald two-sided 95% confidence interval (CI) and p-values for all four treatment comparisons (*Table 2.2.1.1, 2.2.1.2, 2.2.1.3,* and *2.2.1.4* for the efficacy estimand, the attributable estimand, the treatment policy estimand, and the per-protocol estimand, respectively).

Time to first moderate or severe COPD exacerbation will be displayed graphically for each treatment group using a Kaplan-Meier curve (*Figure 2.2.1.1, 2.2.1.2, 2.2.1.3,* and *2.2.1.4* for the efficacy estimand, the attributable estimand, the treatment policy estimand, and the per-protocol estimand, respectively) and analyzed using a log-rank test to compare the curves between the treatments as a supportive analysis. The methodology for how to estimate a Kaplan-Meier curve for the attributable estimand is given in the Details Appendix (SAP Appendix 6). Subjects who do not experience a COPD exacerbation and completed the study (and study treatment) will be censored at Week 52 (Visit 14) for all estimands. For the efficacy and per-protocol estimands, subjects who do not experience a COPD exacerbation and discontinue treatment early will be censored at the date of treatment discontinuation. For the treatment policy estimand, subjects who do not experience a COPD exacerbation and withdraw from the study will be censored at the date of the last assessment or contact (including telephone contact); subjects who withdraw early from the study prior to the Week 52 Visit will be additionally censored at the date that is 52 Weeks after the first dose of study medication. For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin for the hazard ratio of 1.1 will be employed for the upper bound of the two-sided 95% confidence interval for the hazard ratio (for BGF 160 / BFF).

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. Details of the methods may be found in Section 6.4.4.2.

Table 3 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 already-stated negative-binomial analysis framework for the rate of moderate or severe COPD exacerbations. A complete dataset will be created analogously to that described for

the tipping point analysis of Rate of Moderate or Severe Exacerbations 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 1 year. 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 Transition Dyspnea Index

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 9.

At Randomization (Visit 4), the severity of dyspnea at baseline will be assessed using the BDI. At subsequent visits (as per Schedule of Events: see the Schedule of Events in the study protocol), change from baseline will be assessed using the TDI. Scoring and handling of missing items will be conducted in accordance with the user's guide for the TDI score.

The TDI Focal Score will be analyzed using a linear repeated measures (RM) analysis of covariance (ANCOVA) model will include treatment, visit, and the treatment by visit interaction, and ICS use at screening as categorical covariates and log baseline blood eosinophil count, BDI, baseline post-bronchodilator percent predicted FEV_1, and percent reversibility to bronchodilator as continuous covariates. An unstructured (UN) correlation matrix will be used to model additional autocorrelation within subject. If this model fails to converge, a heterogeneous Toeplitz (TOEPH) structure will be used to model correlation between time points from the same subject. If both UN and TOEPH models fail to converge, a first-order heterogeneous autoregressive (ARH(1)) structure will be used to model correlation between time points from the same subject. If the UN, TOEPH, and ARH(1) models all fail to converge, then a first-order autoregressive (AR(1)) structure will be used instead. In ARH(1) and AR(1) models, subject will be included as a random effect. Contrasts will be used to obtain estimates of the treatment differences over the entire 52 weeks, and over 24 weeks. Two-sided p-values and point estimates with two-sided 95% CIs will be produced for each treatment difference (*Table* and *Figure 2.6.1* for the efficacy estimand). For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin of -0.75 will be employed for the lower bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF).

Efficacy data obtained during unscheduled visits will not be used for this analysis.
The efficacy estimand using on-treatment data will be considered the main analysis (*Table* and *Figure 2.6.1*). Analyses of the attributable estimand, the treatment policy estimand, and the per-protocol estimand will be conducted as supportive (*Tables* and *Figures 2.6.2-2.6.4*, respectively).

The attributable estimand (for the analysis over 24 weeks for the TDI focal 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 at Week 24 as described in Section 6.4.7.1.

The TDI focal score over 24 weeks is a secondary endpoint for the ex-US approach. The TDI focal score over 52 weeks is an "other" endpoint. As additional supportive analyses (i.e. other efficacy endpoints), the difference between treatments at each of the individual visits (using the same RM model) will also be evaluated and summarized.

Furthermore, as supportive analyses (i.e. as "other" endpoints), responder analyses will be performed (for the TDI focal score) where responders are defined as a response of 1.0 points or more (corresponding to at least a minor improvement) on average over 24 weeks and on average over 52 weeks. Logistic regression with PROC GENMOD will be used to compare the treatment groups with BDI and log baseline blood eosinophil count, baseline post-bronchodilator percent predicted FEV_1, and percent reversibility to bronchodilator 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 of the 4 treatment comparisons (*Table 2.7.1* for the efficacy estimand over 24 weeks).

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 over 24 weeks, subjects will be considered non-responders if the subject discontinued prematurely from study medication at any time prior to the Week 24 visit. For the TDI responder analyses over 52 weeks, subjects will be considered non-responders if the subject discontinued prematurely from study medication at any time.

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

6.4.5.3 Rescue Ventolin HFA Use

The number of puffs of rescue medication use (i.e. albuterol sulfate or locally available equivalent product) taken in the previous 12 hours will be recorded in the subject diary in the morning and evening. Diary data recorded during the last 7 days of the Screening Period will be used to calculate the baseline. The mean daily number of puffs of rescue medication use (i.e. albuterol sulfate or locally available equivalent product) will be calculated overall and for each of the 4-week intervals (Interval 1 – Interval 13) during the Treatment Period as follows.

The number of puffs of rescue Ventolin HFA taken since the previous (AM or PM) dose will be recorded in the subject diary. For rescue Ventolin HFA use, for every period of time for which the mean number of puffs of rescue Ventolin HFA will be calculated, missing values will be ignored in both the numerator and denominator. 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) over a time interval is then the sum of M_DT and M_DN.

The difference between treatment groups in the change from baseline in average daily rescue medication (i.e. albuterol sulfate or locally available equivalent product) use over 24 weeks (as a secondary endpoint) and over 52 weeks (as an "other" endpoint) will be evaluated. Change from baseline in rescue Ventolin HFA use will be analyzed using a linear RM ANCOVA model will include treatment, 4-week time interval (Interval 1- Interval 13), the treatment by time interval interaction, and screening ICS use as categorical covariates and baseline post-bronchodilator percent predicted FEV_1, baseline rescue Ventolin HFA use, log baseline blood eosinophil count, and percent reversibility to bronchodilator as continuous covariates. An unstructured correlation matrix (UN) will be used to model additional autocorrelation within subject. If this model fails to converge, a TOEPH structure will be used to model correlation between time intervals from the same subject. If both UN and TOEPH models fail to converge, an ARH(1) structure will be used to model correlation between time points from the same subject. If the UN and TOEPH and ARH(1) models all fail to converge, then an AR(1) structure will be used instead. In ARH(1) and AR(1) models, subject will be included as a random effect. Contrasts will be used to obtain estimates of the treatment differences over 24 weeks and the entire 52 weeks. Two-sided p-values and point estimates with two-sided 95% CIs will be produced for each of the 4 treatment differences (*Table* and *Figure 2.10.1.1*). The main analysis will be conducted using the efficacy estimand restricted to the Rescue Ventolin User Population. A sensitivity analysis will use the attributable estimand over 24 weeks (also restricted to the RVU population). Supportive analyses will use the treatment policy estimand, and the per-protocol estimand (*Table* and *Figure 2.10.1.2, 2.10.1.3,* and *2.10.1.4*) restricted to the Rescue Ventolin User Population in each case. Other supportive analyses will use the efficacy estimand not restricted to the Rescue Ventolin User Population (*Table 2.10.1.5*). 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 nighttime rescue medication use (i.e. albuterol sulfate or locally available equivalent product) 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 2.10.2 and 2.10.3* for the efficacy estimand*)*.

The number of relevant respective 4-week interval (1-13) will be used as a categorical covariate in the model. If model convergence is not attained using the 4-week intervals, then 13-week intervals will be used. (In such case, "over 26 weeks" would be summarized instead of "over 24 weeks"). For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg to BFF, a non-inferiority margin of 0.75 puffs per day will be employed for the upper bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF).

The attributable estimand (for the analysis of average daily rescue Ventolin HFA use over 24 weeks) 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 24 as described in Section 6.4.7.1, except that the 95th percentile of the reference arms' distribution will be used for the derivation rather than the 5th percentile.

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 24 weeks (*Table 2.10.1.1a* for the efficacy estimand). 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 have less than that level of change) will be plotted on the Y axis (*Figure 2.10.1.1a* for the efficacy estimand). The word "less than" is used because decreases in rescue are improvements. Subjects who discontinue study medication at any time prior to the Week 24 visit will be considered non-responders. 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 analyses for the attributable estimand will not be performed because such an analysis is not well defined.

Tipping Point Analyses for Rescue Ventolin HFA Use

Robustness of results to missing data will be explored using tipping point analyses for Rescue Ventolin HFA Over 24 Weeks (Ratitch 2013) for the RVU Population. A brief overview of the approach is summarized in the table below. Details of the methods may be found in Sections 6.4.7.8 and 6.4.4.2 and in the Details Appendix to this SAP (Appendix 6).

Table 4 Sensitivity Analyses for Rescue Ventolin HFA Use

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 number of puffs in the treatment arm increased by up to 4 puffs per day 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 number of puffs in the treatment arm increased by up to 4 puffs per day until the p-value ≥ 0.05.	Tipping point analysis: MI using the 95th percentile of the reference arms' 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 number of puffs in the treatment arm increased by up to 4 puffs per day 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 number of puffs in the treatment arm increased by up to 4 puffs per day until the p-value ≥ 0.05.

MNAR = Missing not at random. The tipping point will be shown to at least a precision of 0.1 puffs/day. Imputed values may not be impossible values – i.e. a negative number of puffs of rescue Ventolin HFA. Thus, the values will be imputed from a truncated distribution.

6.4.5.4 St. George's Respiratory Questionnaire

The SGRQ will be used to provide the health status/health-related QoL measurements in this study (see Protocol Appendix 8). 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. A score will be calculated for each component and a "Total" score will be calculated (*Listing 6.1.8*). 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. Missing SGRQ total scores will not be imputed.

The difference between treatment groups in the change from baseline in SGRQ over 24 weeks (as a secondary endpoint) and over 52 weeks (as an "other" endpoint) of treatment will be evaluated using a similar RM approach as for TDI, but using baseline SGRQ score instead of the BDI. Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference. For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin of 3.0 will be employed for the upper bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF).

The main analysis of the SGRQ will use the efficacy estimand. A secondary analysis will use the attributable estimand and supportive analyses will use the treatment policy and the per-protocol estimand. SGRQ total score over 24 weeks is a secondary endpoint for ex-US. SGRQ total score over 52 weeks is an "other" endpoint. As additional supportive analyses (i.e. as other efficacy endpoints), the difference between treatments at each of the individual visits will also be evaluated and summarized (*Tables* and *Figures 2.11.1.1 to 2.11.1.4*)..

Furthermore, responder analyses will be performed where responders are defined as subjects with an improvement (i.e. a decrease in the total SGRQ score) of ≥ 4.0 points at Week 24 (as a secondary endpoint). Responder analyses will be performed also where responders are defined as subjects with an improvement of ≥ 4.0 points at Week 52 and on average over the 52 weeks and on average over 24 weeks as "other" endpoints. For responder analyses at Week 24 (for the Efficacy Estimand and Per-Protocol Estimand) and over 24 weeks (for the Efficacy Estimand), subjects will be considered non-responders if the subject discontinued prematurely from study medication at any time prior to the Week 24 visit. For responder analyses for the Efficacy Estimand at Week 52 and over 52 weeks, subjects will be considered non-responders if the subject discontinued prematurely from study medication at any time. No such imputation of non-response will be made for responder analyses for the Attributable Estimand because the Attributable Estimand has no missing data per se. For responder analyses at Week 24 for the Treatment Policy Estimand, subjects will be considered non-responders if the subject withdrew from the study early at any time prior to the Week 24 visit. Logistic regression will be used to compare the treatment groups with baseline SGRQ Score, log baseline blood eosinophil count, and baseline post-bronchodilator percent predicted FEV_1 and percent reversibility to bronchodilator 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 2.11.3.1* for the efficacy estimand, *Tables 2.11.3.2* for the attributable estimand, *Table 2.11.3.3* for the treatment policy estimand, and *Table2.11.3.4* for the per-protocol estimand, respectively, at Week 24). For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin of 10 percentage points will be employed

for the upper bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF) in percentage response.

The attributable estimand (for the analysis of the change from baseline in total SGRQ score over 24 weeks) 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 24 as described in Section 6.4.7.1, except that the 95th percentile of the reference arms' distribution will be used for the derivation rather than the 5th percentile. For the SGRQ, scoring and handling of missing items will be conducted in accordance with the user's guide for the SGRQ.

Additional sensitivity analyses will be implemented based on a cumulative responder approach as described in Farrar 2006 for the change from baseline in SGRQ score over 24 weeks (*Table 2.11.1.1b* for the efficacy estimand). A cumulative distribution plot by treatment arm (Farrar et al., 2006) will also be produced. The observed change from baseline in SGRQ over 24 weeks will be plotted on the X axis, while the proportion of responders (subjects that equal or have less than that level of change) will be plotted on the Y axis (*Figure 2.11.1.1b* for the efficacy estimand). The word "less" is used because negative changes are improvements. For cumulative responder analyses for the Efficacy Estimand over 24 weeks, subjects will be considered non-responders if the subject discontinued prematurely from study medication at any time prior to the Week 24 visit. The cumulative responder curves for each treatment will then be compared pairwise using Kolmogorov-Smirnov tests. Cumulative responder analyses for the attributable estimand will not be performed because it is not well defined.

<u>Tipping Point Analyses for Percentage of Subjects achieving an MCID of 4 Units or More in SGRQ Total Score at Week 24</u>

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. Details of the methods may be found in Sections 6.4.7.8 and 6.4.4.2 and in the Details Appendix to this SAP (Appendix 6).

Table 5 Sensitivity Analyses for Percentage of Subjects Achieving an MCID of 4 Units or More in SGRQ Total Score at Week 24

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 SGRQ total score in the treatment arm increased by a maximum of 16 units 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 SGRQ total score in the treatment arm increased by a maximum of 16 units until the p-value ≥ 0.05.	Tipping point analysis: MI using the 95th percentile of the reference arms' distribution for the SGRQ Total Score 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 SGRQ total score in the treatment arm increased by a maximum of 16 units 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 SGRQ total score in the treatment arm increased by a maximum of 16 units until the p-value ≥ 0.05.

MNAR = Missing not at random. Tipping point analysis for the treatment policy estimand will not be done for SGRQ Total Score Over 24 Weeks. The tipping point will be shown to at least a precision of 0.01 units. Imputed values may not be impossible values – i.e. a negative SGRQ total score. Thus, the values will be imputed from a truncated distribution.

The multiple imputation will be applied to the continuous total SGRQ scores within the already-stated repeated-measures analysis framework for total SGRQ score. Missing values will first be imputed for the missing total SGRQ scores prior to the computation of whether the subject has attained the MCID (for the sensitivity analysis). The analysis using the imputed data will proceed using logistic regression as described above, followed by a combining of results across the multiple imputations using the formulae of Rubin [Rubin, 1987]

6.4.5.5 EXACT Total Score

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, and the 11-item E-RS total score will be calculated over each 4-week interval of the 52-week Treatment Period. Negative changes of EXACT Total Scores are considered an improvement. The EXACT Total score ranges from 0 to 100. The last 7 days of the Screening Period will be used to calculate the baseline. The mean change from baseline in Exact Total Score (*Tables* and *Figures 2.12.1.1* to *2.12.1.4* for the efficacy estimand, attributable estimand, treatment policy estimand, and per-protocol estimand), RS-Total Score (*Table* and *Figure 2.12.2.1* 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 52 weeks and over 24 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 (1-13) will be used as a categorical covariate in the model. The EXACT Total score over 52 weeks is a secondary efficacy endpoint. The EXACT total score over 24 weeks will be analyzed as an "other" endpoint. The RS-Total score is an "other" endpoint. If model convergence is not attained using the 4-week intervals, then 13-week intervals will be used instead of 4-week intervals. In such case, "over 26 weeks" would be summarized instead of "over 24 weeks".

For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF for the EXACT Total Score, a non-inferiority margin of 1.5 will be employed for the upper bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF).

The attributable estimand (for the analysis of EXACT total score over 52 weeks) 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 24 as described in Section 6.4.7.1, except that the 95th percentile of the reference arms' distribution will be used for the derivation rather than the 5th percentile.

Additional sensitivity analyses will be implemented based on a cumulative responder approach as described in Farrar 2006 for the change from baseline in daily EXACT Total Score over 52 weeks *(Table 2.12.1.1b* for the efficacy estimand). A cumulative distribution plot by treatment arm (Farrar et al., 2006) will also be produced. The observed change from baseline in mean daily EXACT Total score over 52 weeks will be plotted on the X axis, while the proportion of responders (subjects that equal or have less than that level of change) will be plotted on the Y axis (*Figure 2.12.1.1b* for the efficacy estimand). For cumulative responder analyses over 52 weeks, subjects will be considered non-responders if the subject discontinued prematurely from study medication at any time. 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 analyses for the attributable estimand will not be performed because such an analysis is not well defined.

6.4.5.6 Time to Death: All Cause, Respiratory

Time to death (All Cause) in weeks will be summarized using a Kaplan-Meier curve. For these endpoints, the primary population will be the treatment policy estimand. The event time for a death will be the date of death. Subjects will be censored at the date of last contact. The time to death will be used to compare treatments using a Cox regression model, adjusted for baseline post-bronchodilator percent-predicted FEV_1 and baseline age as covariates. Hazard ratios with Wald 2-sided 95% CIs for these ratios will also be provided for all four treatment comparisons (*Table* and *Figure 2.3.1* for the treatment policy estimand). Separate analyses will be done for time to death from all causes and time to death from respiratory causes. A log rank test will be performed as a sensitivity analysis. Pearl and the adjudication committee will agree on search terms (based on the prevailing version of the MedDRA dictionary) to identify mortalities due to possible respiratory causes. Only those deaths identified as being due to respiratory causes by the adjudication committee will be considered as events for this analysis. For the analysis of all-cause death, subjects who do not die will be censored at the day of last contact. For subjects who complete the study, the date of death will be censored at the Week 52 visit date. For subjects who withdrew early from the study, the date of death will be censored at the date that is 52 Weeks after the date of first study medication.

As an additional supportive analysis (i.e. other efficacy endpoint) time to death (respiratory) will be analyzed in the same manner as time to death (all cause) and will be tabulated in *Table and Figure 2.4.1* for the treatment policy estimand. Respiratory deaths will be determined from the Adjudication CRF page. For the analysis of respiratory deaths, subjects who do not die from respiratory causes will be censored according to the censoring rules stated above for time to all-cause death. If there are fewer than 30 such deaths, then only descriptive statistics will be presented.

6.4.5.7 Rate of Severe COPD Exacerbations

The rate of severe COPD exacerbations (*Tables 2.1.1.10, 2.1.1.10d, and 2.1.1.10j* for the efficacy, attributable, and treatment policy estimands) will be analyzed in a manner similar to the primary efficacy variable, the rate of moderate or severe COPD exacerbations.

The logarithm of time at risk of experiencing a moderate or severe (see explanation below) COPD exacerbation will be used as an offset variable in the model for the rate of severe COPD exacerbations. Time during a severe exacerbation or in the 7 days following a severe exacerbation will not be included in the calculation of exposure (i.e. time at risk). However, the start day of a severe COPD exacerbation will not be excluded from the time at risk.

Sensitivity analyses for missing data will be identical to those described in Section 6.4.4.2 *(Tables 2.1.1.10b, 2.1.1.10c, 2.1.1.10e, and 2.1.1.10h).*

6.4.6 Analysis of Other Endpoints

Analyses of other endpoints will be performed for the efficacy estimand only.

6.4.6.1 Time to First COPD Exacerbation of Any Severity and Time to First Severe COPD Exacerbation

Time to first COPD exacerbation of any severity and time to first severe COPD exacerbation will be analyzed in a manner similar to the time to first moderate or severe COPD exacerbation as defined under secondary variables. However, sensitivity analyses will not be performed for these endpoints.

The severities of exacerbations and their onset dates were defined previously. Time to first COPD exacerbation of any severity is the time from first dose of study medication (or from randomization for any subjects randomized but not treated) to the time of onset of the first COPD exacerbation (mild, moderate, or severe). Time to first severe exacerbation is the time from first dose of study medication (or from randomization for any subjects randomized but not treated) to the time of onset of the first severe COPD exacerbation.

Estimated adjusted hazard ratios for each endpoint will be displayed along with the associated Wald two-sided 95% CI and p-values *(Tables 2.2.2* and *2.2.3* for the efficacy estimand for exacerbations of any severity and for severe exacerbations). Kaplan-Meier curves for each endpoint will be presented in *Figures 2.2.2* and *2.2.3* for exacerbations of any severity and for severe exacerbations for the efficacy estimand.

6.4.6.2 Rate of COPD Exacerbations of Any Severity

The rate of COPD exacerbations of any severity *(Tables 2.1.1.7* for the efficacy estimand) will be analyzed in a manner similar to the primary efficacy variable, the rate of moderate or severe COPD exacerbations.

The logarithm of the time at risk of experiencing an exacerbation will be used as an offset variable in the model for the rate of COPD exacerbations of any severity. 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). However, the start day of a COPD exacerbation will not be excluded from the time at risk. Data related to COPD exacerbations of any severity are listed in *Listings 6.1.2.1, 6.1.2.2,* and *6.1.2.3*. 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 Treatment Failure

Treatment failure is defined as a moderate or severe COPD exacerbation, 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 2.5.1*). Subjects who do not experience a treatment failure will be censored at their Week 52 Visit date. Subjects will also be censored at the at-risk end date for moderate or severe COPD exacerbations. 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 COPD exacerbation history, log baseline blood eosinophil count, region (US & Canada, Asia, Western Europe, Eastern Europe, Mexico & Central America & South America, Australia & New Zealand & South Africa), and ICS use at screening. Estimated adjusted hazard ratios will be displayed along with associated 95% CI and p-values (*Table 2.5.1*).

6.4.6.4 Percentage of Days with "No Rescue Ventolin HFA Use" Over the Treatment Period

As a supportive analysis, percentage of days with "no rescue Ventolin HFA use" over 52 weeks will be analyzed. A "day with no rescue use" is defined as any day where the subject reported having taken zero puffs of rescue Ventolin HFA. Percentage of days with no rescue use = 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 percentage of days with "no rescue use" will be summarized by treatment and analyzed using ANCOVA with baseline average daily rescue Ventolin HFA use, log baseline blood eosinophil count, post-bronchodilator percent predicted FEV_1, and percent reversibility to bronchodilator as continuous covariates and ICS use at screening as a categorical covariate. Two-sided p-values and point estimates with 2-sided 95% CIs will be produced for each treatment difference (*Table 2.10.4* for the efficacy estimand).

For the efficacy estimand, the analyses of Percentage of Days with "No Rescue Ventolin HFA Use" will be restricted to the Rescue Ventolin User Population.

6.4.6.5 European Quality-of-Life-5 Dimension-5 Level Questionnaire

The European Quality-of-Life-5 Dimension-5 Level Questionnaire (EQ-5D) 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 End of Treatment (%, 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 2.13.1* for the efficacy estimand). Descriptive statistics for the index score

(*Table 2.13.2* for the efficacy estimand) and VAS (*Table 2.13.3* for the efficacy estimand) will be presented by treatment group. VAS scores over 52 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 2.13.2* for the index score, *Table* and *figure 2.13.3* for VAS for the treatment policy estimand and *Table* and *figure 2.13.3.1* for VAS for the efficacy estimand). EQ-5D-5L data are listed in *Listing 6.1.9*.

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).

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

6.4.7 Analysis of PFT Sub-Study

Pulmonary function testing is planned to be conducted in a subset of approximately 3060 subjects at the randomization visit and at Day 1, Weeks 4, 12, 24, 36, and 52. Spirometry data are listed in *Listings 6.2* and *6.3.1*. Participation in the PFT sub-study is listed in *Listing 9.7*.

The efficacy, attributable, treatment policy, and per-protocol estimands will be analyzed in an analogous manner to that described for the secondary endpoints in Section 6.4.5. The efficacy estimand will use all observed data in the mITT Population. The treatment policy estimand will use all observed data in the ITT Population. The attributable estimand will use the mITT Population but then impute missing post-treatment discontinuation data using the observed data model); further, if the reason for discontinuation is attributable to lack of efficacy or tolerability then a penalty will be applied to the imputed value. All other missing data will be imputed using the observed data model. More detail about the attributable estimand is given in Section 6.4.1 and in the Details Appendix to the SAP. The per-protocol estimand will use the PP Population.

6.4.7.1 Change from Baseline in Morning Pre-dose Trough FEV_1

Change from baseline in morning pre-dose trough FEV_1 at each visit is defined as the mean 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.

The change from baseline in morning pre-dose trough FEV_1 will be analyzed using an RM linear mixed model. The RM linear mixed model will include baseline FEV_1, log baseline blood eosinophil count, and percent reversibility to bronchodilator as continuous covariates and visit, treatment, the treatment by visit interaction, and ICS use at screening as categorical covariates. Baseline FEV_1 and baseline eosinophil count are defined in Section 6.4.2. Pre-dose trough is defined as the mean of the non-missing -60 minute and -30 minute values obtained prior to dosing at the respective post-baseline visit. An unstructured correlation matrix (UN) will be used

to model additional autocorrelation within subject. If this model fails to converge, a TOEPH structure will be used to model correlation between time points from the same subject. If both UN and TOEPH models fail to converge, an ARH (1) structure will be used to model correlation between time points from the same subject. If the UN, the TOEPH, and the ARH (1) models all fail to converge, then an AR (1) structure will be used. In ARH (1) and AR (1) models, subject will be included as a random effect. Contrasts will be used to obtain estimates of the treatment differences at Week 24, over 24 weeks, over Weeks 12 to 24, over 52 weeks (as an "other" endpoint), as well as at each post-randomization visit (as "other" endpoints). Two-sided p-values and point estimates with two-sided 95% CIs will be produced for each treatment difference (*Table* and *Figure 2.14.1* to *2.14.4* for the efficacy estimand, the attributable estimand, the treatment policy estimand, and per-protocol estimand, respectively). For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin of -50 ml will be employed for the lower bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF).

For the attributable estimand (for the analysis at Week 24 and the analysis over 24 weeks), data that are missing due to treatment discontinuation will be imputed based on von Hippel's maximum likelihood approach. If the reason is reasonably attributable to tolerability or lack of efficacy. Multiple imputation for missing values for morning pre-dose trough FEV_1 and FEV_1 AUC_{0-4} will use mean changes from baseline based on von Hippel's maximum likelihood approach. When missingness is reasonably attributable to treatment (see Section 6.4) as described in the Details Appendix (Appendix 6). Other missing data are to be imputed using the observed data model.

Exploration of the robustness of findings to missing data are discussed below in Section 6.4.7.8 .

Change from baseline in morning pre-dose trough FEV_1 at each visit is defined as the mean 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.

6.4.7.2 **FEV_1 AUC_{0-4} at Week 24 (US) and Over 24 Weeks (ex-US)**

FEV_1 AUC_{0-4} will be calculated using the trapezoidal rule, after first having subtracted the baseline FEV_1 value, and the AUC will be transformed into a weighted average by dividing by the time in hours from dosing to the last measurement included (typically 4 hours). For the efficacy estimand, the treatment policy estimand, and the per-protocol estimand, only one non-missing post-dose value is required for the calculation of AUC; for the per-protocol estimand, AUC will be calculated provided that there are at least 2 non-missing post-dose values. Actual time from dosing will be used if available; otherwise scheduled time will be used. The differences between treatment groups in FEV_1 AUC_{0-4} will be evaluated using an RM linear mixed ANCOVA model (similar to that used for pre-dose trough FEV_1) with baseline FEV_1, log baseline blood eosinophil count, and percent reversibility to bronchodilator as continuous covariates and visit, treatment, the treatment by visit interaction, and ICS use at screening as

categorical covariates. Contrasts will be used to obtain estimates of the treatment differences at Week 24, over 24 weeks, over Weeks 12 to 24, over 52 weeks (as an "other" endpoint), as well as at each post-randomization visit (as "other" endpoints). Two-sided p-values and point estimates with two-sided 95% CIs will be produced for each treatment difference (*Table* and *Figure 2.16.1* to *2.16.4* for the efficacy estimand, the attributable estimand, the treatment policy estimand, and the per-protocol estimand, respectively). For the non-inferiority comparison of BGF MDI 160/14.4/9.6 µg MDI to BFF, a non-inferiority margin of -75 ml will be employed for the lower bound of the two sided 95% CI for the treatment difference (BGF 160 minus BFF).

The attributable estimand (for the analysis at Week 24 and the analysis over 24 weeks) will be computed in the same manner as it is for Change from Baseline in Morning Pre-dose Trough FEV_1 at Week 24 as described in Section 6.4.6.1.

Exploration of the robustness of findings to missing data are discussed below in Section 6.4.7.8 and are discussed in the Details Appendix to this SAP (Appendix 6).

6.4.7.3 Other PFT Sub-Study Analyses

In addition to the two primary PFT endpoints discussed above (Sections 6.4.7.1 and 6.4.7.2), the following analyses will be carried out.

Peak change from baseline in FEV_1 (assessed within a visit) over 24 weeks, over Weeks 12 to 24, and over 52 weeks, and at each post-randomization visit where measured will be estimated and compared between treatment groups using a linear mixed RM model with the same model as pre-dose trough FEV_1 (*Table* and *Figure 2.17.1* for the efficacy estimand).

6.4.7.4 Time to Onset of Action Assessed Using FEV1 on Day 1.

The onset of action will be determined for each treatment using the post-dosing FEV_1 assessments from Day 1. The onset of action for each product (BGF MDI 320, BGF MDI 160, GFF MDI, and BFF MDI) will be defined as the first time point where the mean change from baseline exceeds 100 mL. Supportive analyses may be conducted using alternative definitions of onset of action. The resulting table and figure are *Table* and *Figure 2.29.1* for the efficacy estimand.

6.4.7.5 Rate of decline in pre-dose FEV1 and FEV1 AUC0-4 over 52 weeks

The rate of decline in pre-dose FEV_1 will be analyzed as follows. Pre-dose trough FEV_1 will be analyzed with a linear mixed model with random subject slopes of pre-dose trough FEV_1 versus time and random subject intercepts. The time is calculated as weeks between the date of the first dose of study treatment and the date of the visit, that is (Date of Visit – Date of First Dose +1)/7. Only post-baseline visits will be used in the repeated measures model. An unstructured correlation matrix will be employed for these two random effects. The model will fit a distinct fixed mean for each treatment and a distinct fixed slope (of FEV_1 versus time) for each treatment. The model will include the following categorical variables: treatment, smoking status at screening, and ICS use at screening. The model will include the following continuous

variables: baseline FEV_1, percentage reversibility to bronchodilator and log baseline blood eosinophil. The following interactions will also be included in the model: the interaction between treatment and time, and the interaction between screening smoking status and time and the interaction between baseline FEV_1 and time. The two categories of smoking status for the analysis are current smoker and former smoker. If there are any non-smokers, they will be grouped together with the former smokers for this analysis. The rate of decline (the negative of the slope) will be estimated and compared between treatments (*Table 2.27.1* for the efficacy estimand).

The rate of decline in FEV_1 AUC_{0-4} over 52 weeks will be analyzed in a similar manner (*Table 2.28.1* for the efficacy estimand). For AUC_{0-4}, the Day 1 value will be included as a time point in the analysis (as time zero).

6.4.7.6 Endpoint-Specific Evaluability for PFT analyses

Peak FEV_1 will be included in the analyses as long as there is at least one non-missing post-dose value.

6.4.7.7 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, however, for example during a COPD exacerbation unrelated to treatment, extreme and atypical values can arise. Such values have high influence on estimation of variance parameters and on standard errors of fixed effect estimates. The distribution of residuals and influence statistics for morning pre-dose trough FEV_1 will be examined to identify such cases. In the event that a single, or small number of such outlying values, are found to exist and to be highly influential, the effects may be ameliorated by removal of the outlier or by the use of nonparametric methods. These analyses will be conducted if warranted to demonstrate the robustness of the results and reported in the Statistical Methods Appendix to the CSR.

The assumption of normality in the data for morning pre-dose trough FEV_1 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. baseline FEV_1, residuals vs. percent reversibility to bronchodilator at Visit 2, residuals vs. treatment, residuals vs. smoking status (non-smoker/current smoker/former smoker), residuals vs. ICS use (yes/no), and by box plots of residuals for model variables with a potential effect on variance (treatment, visit, smoking status, and ICS use). 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 correlation matrix failed to converge), and/or ICS use categories (yes/no).

6.4.7.8 Sensitivity Analysis of PFT Sub-Study for Missing Data

Sensitivity analyses will be conducted to evaluate the robustness of the primary analysis findings of the PFT sub-study to missing data. The estimands of interest in this study to evaluate the

effect of treatment compared to control (for example BGF MDI 320/14.4/9.6 µg vs. GFF MDI) at Week 24 and over 24 weeks of treatment include an efficacy estimand, an attributable estimand, and a treatment policy estimand. For the efficacy estimand, since the primary analyses of morning pre-dose trough FEV_1 and FEV_1 AUC_{0-4} use maximum likelihood based approaches, they are valid under the MAR assumption (Little and Rubin, 2002). In order to evaluate the robustness of the findings to this assumption, sensitivity analyses will be performed under varying assumptions for data considered likely to be missing not at random (MNAR).

Several types of statistical models have been proposed to analyze clinical study data under such assumptions. The first approach that will be implemented for this study is the use of pattern-mixture models (PMMs). The second approach will implement a cumulative responder analysis.

Secondary analyses of the treatment policy estimand will be conducted in the ITT Population where all observed data will be utilized regardless of whether subjects remain on randomized treatment. A sensitivity analysis will be conducted where missing data in the treatment arm are imputed with the benefit of treatment decremented by up to 500 mL until the p-value ≥ 0.05.

The following table summarizes the multiple imputation-based sensitivity analyses under the PMM framework that will be undertaken,

Table 6 Sensitivity Analyses for Morning Pre-dose Trough FEV_1 and FEV_1 AUC_{0-4}

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 von Hippel's maximum likelihood approach using the 5th percentile of the reference arms' 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) 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. No tipping point analyses will be done for the treatment policy estimand for morning pre-dose Trough FEV_1 over 24 weeks and FEV_1 AUC_{0-4} over 24 weeks. The tipping point will be shown to 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 an efficacy estimand, which 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.

Although the analysis for the attributable estimand starts with the same amount of missingness (as the efficacy estimand has), 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.

Tipping-point analyses will be conducted to examine the impact of varying the treatment mean for missing data in subjects who discontinue BGF MDI. MI techniques will be used to impute the missing data for these patients by varying the mean in the treatment arm. The imputed value in the treatment arm will be decremented by up to 500 mL until the p-value ≥ 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 BGF MDI whose missing data are determined to be MNAR.
>
> Tipping Point #2: this analysis will impute diminished effects for all missing data in the BGF 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 using the 5th percentile of the reference arms' 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) vs. p-values will also be produced. Details of the sensitivity analyses will be discussed in the Statistical Methods Appendix to the CSR.

Additional sensitivity analyses will be implemented based on a cumulative responder approach as described in Farrar 2006 for the change from baseline in morning pre-dose trough FEV_1 and FEV_1 AUC_{0-4} at Week 24 (*Tables 2.15* and 2.16.10 for the efficacy estimand) and over 24 weeks

(*Tables 2.15.1* and *2.16.11* for the efficacy estimand). A cumulative distribution plot by treatment arm (Farrar et al., 2006) will also be produced. The observed change from baseline in each of these lung function variables at Week 24 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 2.15* and *2.16.10* for the efficacy estimand). The observed change from baseline in each of these lung function variables over 24 weeks will also be plotted (*Figures 2.15.1* and *2.16.11).* For cumulative responder analyses at Week 24, subjects will be considered non-responders if the subject discontinued prematurely from the study medication at any time prior to the Week 24 visit. 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. Cumulative responder analyses for the attributable estimand will not be performed because it is not well defined.

6.4.8 Subgroup Analyses for Efficacy

Selected tabulations (named below) will be provided for the following subgroups:

- China
- Asia (Asia is defined by country rather than by race and includes China, Japan, South Korea and Taiwan)
- US
- Europe (including Russia)
- North America
- Other (including Mexico, Central America, South America, Australia, New Zealand, and South Africa)
- Baseline Blood Eosinophil Count (< 150 cells per mm^3 or ≥ 150 cells per mm^3) (Day 1)
- Racial groups:
  - Black or African American
  - White
  - Asian
  - Other, Native Hawaiian or Other Pacific Islander, or American Indian or Alaska Native
- Age groups:
  - Age < 65 years
  - Age ≥ 65
- Sex groups:
  - Male
  - Female

- ICS Use at Screening (for the designated efficacy endpoints only, not safety):
  - Yes
  - No
- COPD Exacerbation History (for analyses of exacerbations only):
  - 1 (not including any values of zero if they exist)
  - ≥ 2
- Post-bronchodilator FEV_1 (< 50%, ≥ 50% Predicted)
- Baseline Blood Eosinophil Count < 150 cells per mm^3 and COPD Exacerbation History < 2
- Baseline Blood Eosinophil Count < 150 cells per mm^3 and COPD Exacerbation History ≥ 2
- Baseline Blood Eosinophil Count ≥ 150 cells per mm^3 and COPD Exacerbation History < 2
- Baseline Blood Eosinophil Count ≥ 150 cells per mm^3 and COPD Exacerbation History ≥2

The distribution of baseline characteristics used in the primary model – ICS use at screening, region, exacerbation history (1, ≥ 2, as well as mean), baseline eosinophil count, post-bronchodilator FEV_1, reversibility to albuterol – will be summarized within subgroups, both overall and by treatment group.

To investigate whether the differences in treatment effects among countries are due to chance, Galbraith plots (Anzures-Cabrera and Higgins 2010), with corresponding tables, and Q-Q (quantile-quantile) plots will be provided. A Galbraith plot is a scatter plot of standardized treatment effect (effect estimates divided by their standard errors), against the precision of the study estimate (expressed as 1/SE, i.e., the inverse standard error). Imprecise estimates of effect lie near the origin and precise estimates further away, so that distance from the origin conveys the relative amounts of information in the different countries. Vertical scatter of points in a Galbraith plot reflects the extent of heterogeneity. After computing the estimates for the Galbraith plot, a simple linear regression of the standardized treatment effect on the inverse standard error will be fit to calculate a least squares regression line for each treatment pair. The regression line will be constrained to include the origin. Lines drawn at a vertical distance of ±2 above and below the regression-through-the-origin line represent an approximate 95% confidence region. Under a fixed-effect meta-analysis model, 95% of countries will, on average, lie between these two lines. Effects out of this band are statistically significant at $p<0.05$ and it is possible to calculate the probability of observing at least one estimate as extreme as this in a particular country after accounting for the number of countries assessed. A Q-Q plot of observed vs. expected standardized residuals from the normal probability distribution will be provided. The Q-Q plot will be used to indicate whether the effects are within the expected distribution of the standardized effects. The straight line is the expected effects – i.e. the line for which the observed standardized residual is equal to the expected value of the respective order statistic –

from the standard Normal distribution and two dotted lines represent an approximate 95% confidence region of the observed effects above or below the straight line of expected effects. The confidence region is based on the theoretical standard normal distribution. The expected z is simply the probit (inverse Gaussian CDF) of the empirical CDF of the set of z-statistics, corresponding to each z-statistic. CDF denotes cumulative distribution function. For the Galbraith plots and the Q-Q plots, countries with fewer than 20 data points for continuous endpoints – or with fewer than 20 events for count endpoints or time-to-event endpoints – will not be included in the plot. In the case of no ties, the empirical CDF is the ordinal number of the z-statistic minus 0.5, all divided by the number of z-statistics.

The following tables and figures will be provided by subgroup (with the exception of COPD exacerbation history

- Rate of Moderate or Severe COPD Exacerbations (*Tables 4.1.1.1 to 4.1.1.21* for the efficacy estimand; Galbraith, and Q-Q Plots for the efficacy estimand: Galbraith *Figures 4.1.1.21* to *4.1.1.24* for the efficacy estimand; Q-Q Plots *4.1.1.25* to *4.1.1.28* for the efficacy estimand)
- Change from baseline in morning pre-dose trough FEV_1 at Week 24 and over 24 weeks (*Tables 4.1.2.1* through *4.1.2.13*)
- FEV_1 AUC_{0-4} at Week 24 and over 24 weeks (*Tables 4.1.3.1* through *4.1.3.13*)

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 primary efficacy endpoint tables for each subgroup (*Tables 4.1.1.1* to *4.1.3.13*).

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 4.2.1* for moderate or severe COPD exacerbations and *4.2.2* for morning pre-dose FEV_1 and FEV_1 AUC_{0-4} for the efficacy estimand). Should any country or region effects be identified, shrinkage estimates may be generated in order to further understand the impact of these effects (Carroll and Fleming, 2013).

Subgroup analyses of the rate of moderate or severe exacerbations are being conducted using a 150 cells/mm^3 cutoff as stated in the beginning of this section. It is acknowledged 150 cells/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 subgroups defined by different cut points as well as potentially using generalized additive models (GAMs) (or LOESS) that combine nonparametric regression for the relationship of eosinophil levels to response with a negative binomial model.

A more comprehensive analysis will be performed for the China and Asia subgroups, to be detailed in a supplementary China SAP, which will serve as a basis for developing a China CSR to support the China NDA filing. Adjustment for multiplicity will not be employed due to the exploratory nature of the China/Asia subpopulation analysis.

6.4.9 Control of Type I Error

All comparisons will be for superiority, with the exception of BGF MDI 160/14.4/9.6 µg MDI to BFF. The comparisons of BGF MDI 160/14.4/9.6 µg to BFF MDI 320/9.6 µg on COPD exacerbations will be for non-inferiority followed by superiority; however, attaining statistical significance in the superiority comparison is not a pre-requisite to proceeding down the testing hierarchy.

If BGF MDI 320/14.4/9.6 µg significantly reduces the rate of moderate or severe COPD exacerbations compared to both GFF MDI 14.4/9.6 µg and BFF MDI 320/9.6 µg (using first the efficacy estimand and then the attributable estimand, which is a secondary endpoint), then primary endpoints from the lung function sub-study will be assessed. For the US approaches, these are first FEV_1 AUC_{0-4} at Week 24 for the comparison of BGF MDI 320/14.4/9.6 µg to BFF MDI 320/9.6 µg and then the change from baseline in morning pre-dose trough FEV_1 at Week 24 for the comparison of BGF MDI 320/14.4/9.6 µg to GFF MDI 14.4/9.6 µg. For ex-US Approaches, these are first FEV_1 AUC_{0-4} over 24 weeks for the comparison of BGF MDI 320/14.4/9.6 µg to BFF MDI 320/9.6 µg and then the change from baseline in morning pre-dose trough FEV_1 over 24 weeks for the comparison of BGF MDI 320/14.4/9.6 µg to GFF MDI 14.4/9.6 µg. If these comparisons are statistically significant, then for both the US and Ex-US approaches, the rate of moderate or severe exacerbations in the baseline exacerbation history ≥ 2 exacerbations in the last year category will be compared between BGF MDI 320/14.4/9.6 µg and GFF MDI 14.4/9.6 µg and between BGF MDI 320/14.4/9.6 µg and BFF MDI 320/9.6 µg, as will the remaining secondary endpoints. BGF MDI 160/14.4/9.6 µg will follow a similar approach, except that the comparison to BFF 320/9.6 µg will be for non-inferiority first. If BGF MDI 160/14.4/9.6 µg significantly reduces the rate of moderate or severe COPD exacerbations compared to GFF MDI 14.4/9.6 µg and is non-inferior to BFF MDI 320/9.6 µg, then primary endpoints from the lung function sub-study will be assessed as outlined above.

For both registration approaches, US and ex-US, and for each dose of BGF MDI, if the primary and first secondary measures are statistically significant for all comparisons, then Type I error for the remainder of the secondary measures with the exception of time to death (all cause) will be controlled by using Hochberg within each comparison [Hochberg 1988]. If all of the secondary measures are significant, then time to death (all cause) will be tested within each comparison.

For the US, the remainder of the secondary endpoints, controlled by the Hochberg procedure, are as follows:

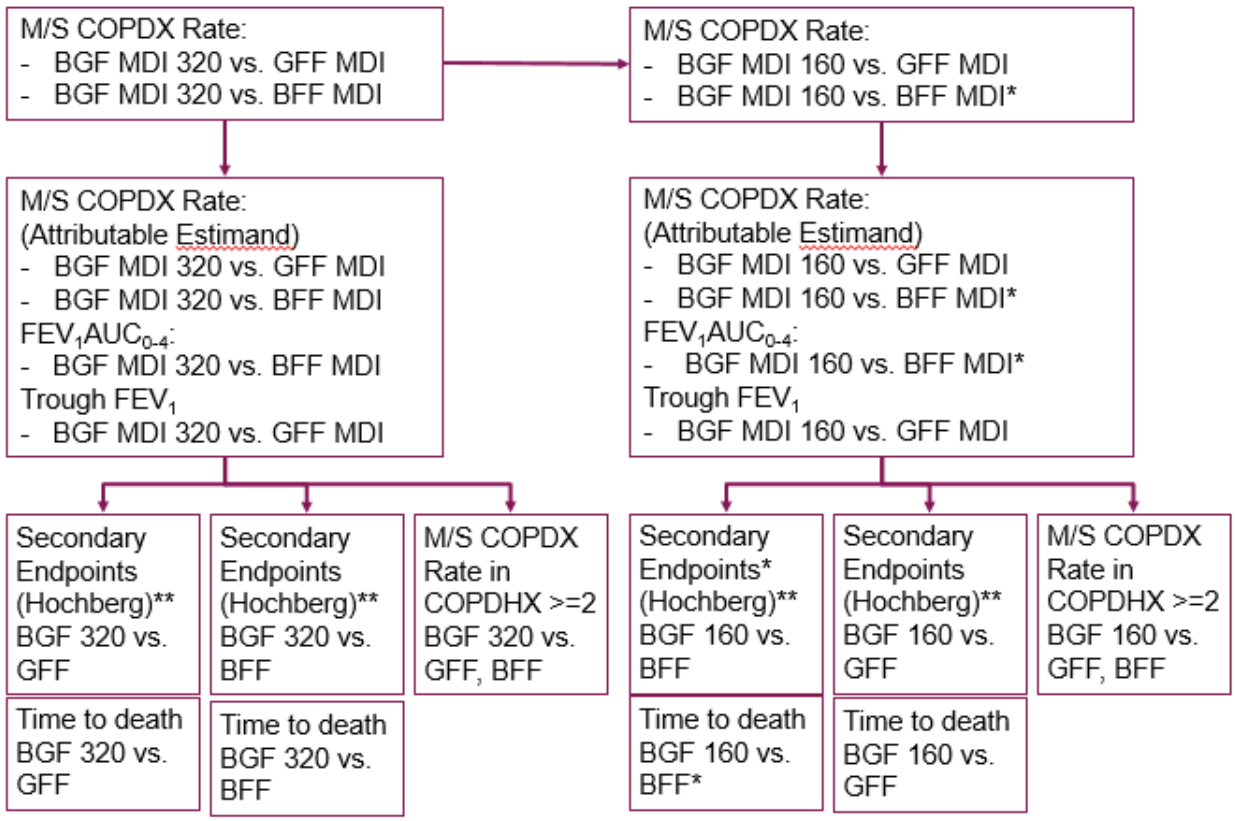
- Time to first moderate or severe COPD exacerbation
- Change from baseline in average daily rescue Ventolin Hydrofluoroalkane (HFA) use over 24 weeks
- Percentage of subjects achieving a minimal clinically important difference (MCID) of 4 units or more in SGRQ total score at Week 24
- Rate of severe COPD exacerbations

For ex-US, the remainder of the secondary endpoints, controlled by the Hochberg procedure, are as follows:

- Time to first moderate or severe COPD exacerbation
- Change from baseline in average daily rescue Ventolin Hydrofluoroalkane (HFA) use over 24 weeks
- Transition Dyspnea Index (TDI) focal score over 24 weeks
- Change from baseline in the Exacerbations of Chronic Pulmonary Disease Tool – (EXACT) total score over 52 weeks
- Change from baseline in St. George's Respiratory Questionnaire (SGRQ) total score over 24 weeks
- Rate of severe COPD exacerbations

The family-wise Type I error rate is maintained throughout the trial, including the interim analysis by one-sided group sequential testing. Further details about maintaining the Type I error throughout multiple tests may be found in Section 3.4.

Figure 2 Order of Hypothesis Testing for Type I Error Control: Ex-US and US Approaches



*All comparisons of BGF160 vs BFF are for non-inferiority using the per-protocol estimand followed by superiority. Superiority is not required to advance to the next comparison. All other comparisons are for superiority and use the efficacy estimand unless otherwise stated.
Endpoints for the US are at Week 24, and for ex-US over 24 weeks wherever applicable (PFT, QoL, Symptoms)
**Hochberg-controlled secondary endpoints: T2f M/S COPDX, Rescue Use, Rate of Severe COPDX, SGRQ (ex-US), SGRQ responders (US), EXACT Total (ex-US), and TDI (ex-US).

After testing the first secondary endpoint, Hochberg's step-up procedure (1988) is performed for the remaining "secondaries" within each treatment comparison using the following steps:
Start by ordering the p-values (from lowest to highest) P (1) … P (m), and let the associated hypotheses be H(1) … H(m). For a given α, let R be the largest k such that P(k) <= α/(m-k+1). Reject the null hypotheses H(1) … H(k).

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 relevant 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 are events with an onset date on or after the first date of dose and up to and including the date of completion of study treatment or the day after the date of premature discontinuation from study treatment. 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 on [or after] the day after the dateof completion of study treatment or on or after the day after the day after (i.e. 2 days after) the date of premature discontinuation of study treatment. The exception is that deaths are still considered to be during the Treatment Period if any adverse event that led to that death started during the Treatment Period. This rule applies to safety analyses.

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, 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 (serious adverse events), and deaths during the post- -treatment period will be tabulated and listed. Tabulations of the incidence of deaths, AEs by SOC, and SAEs by SOC will be provided using information collected after treatment discontinuation (as alternative tables).

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 term will be provided in *Listing 7.2*.

An adverse event is considered on-treatment (i.e. treatment-emergent) if an event occurs after the first dose of 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 study medication plus one day or on or before the date of completion of study medication. An adverse event that begins on the same date as the first dose of study medication is treatment-emergent if the AE begins after the time of first dose or if the time of AE onset is unknown. Adverse events with onset after the date of premature discontinuation from study treatment plus one day or after the date of

completion of study treatment will not be considered treatment-emergent, but will be tabulated separately (*Table 3.2.1.2*) and listed in adverse event data listings (*Listing* 7.1). Adverse events that occur between the time the subject signs the informed consent form (ICF) 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. Additionally, if an AE has an onset date during treatment and has an outcome of death, that death will be considered on treatment even if the date of death is after the last date of treatment+1.

The incidence of an AE will be defined as the number and percentage of subjects experiencing an event. Incidence of treatment-emergent adverse events (TEAEs) (*Tables 3.2.1.1* and *3.2.1.2*), serious TEAEs (*Table 3.7.1*), major adverse cardiovascular events (MACE) or pneumonia (*Table 3.13.3*), neoplasms (*Tables 3.10.1, 3.10.2*), TEAEs leading to study drug discontinuation (*Table 3.5*), and adverse events leading to death by adjudicated category (*Tables 3.15.1.3* and *3.15.1.4*) will be summarized by treatment group. They will be tabulated at the level of the MedDRA preferred term and the MedDRA system organ class (SOC). No hypothesis tests will be performed. Incidence of TEAEs and incidence of MACE or pneumonia will also be tabulated by the following time intervals: 0 to <= 24 Weeks and > 24 Weeks. Adverse events leading to death will also be summarized by adjudicated category (*Tables 3.15.1.1* and *3.15.1.2).*

An overview table will be prepared with the incidence of subjects with at least one 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 early withdrawal, and a report of death (*Table 3.1*). Overview tables will also be prepared by the following time intervals: 0 to <= 24 Weeks and > 24 Weeks.

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*.

The listing of adverse events will provide the severity, maximum severity, relationship to study drug as assessed by the investigator, action taken and outcome for each adverse event. Any SAEs reported will be listed (*Table 3.8.1).* Adverse events leading to discontinuation of study medication will be listed *(Table 3.6)*. A listing of any reported deaths during the study (prior to randomization, during treatment, or during the post-treatment-discontinuation follow-up) will be provided (*Table 3.15.2*); 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.

The following is a list of summary tabulations that 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:

1. The incidence of all treatment-emergent adverse events *(Table 3.2.1.1)*
2. The incidence of all treatment-emergent adverse events for the following time intervals: 0 to <= 24 Weeks and > 24 Weeks) *(Tables 3.2.1.1.1 and 3.2.1.1.2)*
3. The incidence of treatment-emergent adverse events with onset during the post-treatment period (*Table 3.2.1.2*)
4. A summary tabulation will also be prepared for the incidence of treatment-emergent adverse events occurring in at least 2% of subjects in any treatment (*Table 3.2.2,* sorted by descending frequency of events in a preferred term).
5. The incidence of treatment-emergent adverse events occurring in SMQs (Standard MedDRA Queries)/groupings of interest (*Table 3.2.3*)
6. The incidence of non-serious treatment-emergent adverse events occurring in >= 5% of subjects in a treatment (*Table 3.2.4*)
7. The incidence of all treatment-emergent adverse events suspected to be drug-related (*Table 3.4*)
8. The incidence of discontinuation from study treatment due to a treatment-emergent adverse event (*Table 3.5*)
9. The incidence of treatment-emergent serious adverse events (*Table 3.7.1*)
10. The incidence of treatment-emergent serious adverse events with onset during the post-treatment period (*Table 3.7.2*)
11. The incidence of all treatment-emergent serious adverse events suspected to be drug-related (*Table 3.9*)
12. The incidence of all neoplasms (all cancer) (*Table 3.10.1)*
13. The incidence of all neoplasms (excluding non-melanoma skin neoplasms) (*Table 3.10.2)*
14. The incidence of all treatment-emergent adverse events by highest severity to treatment (*Table 3.11.1* through *3.11.4* for the four treatments)
15. The incidence of adjudicated MACE or pneumonia adverse events (*Table 3.13.3)*
16. The incidence of adjudicated MACE or pneumonia adverse events for the following time intervals: 0 to <= 24 Weeks and > 24 Weeks) (*Table 3.13.4 and 3.13.5)*
17. The incidence of adverse events leading to death (*Table 3.15.1.3)*
18. The incidence of adverse events leading to death with onset during the post-treatment period (*Table 3.15.1.4)*
19. 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 (*Table 3.3*)
- Frequency and rate of SAEs (Table 3.8.3)
- Frequency and rate of neoplasms (Tables 3.10.3 and 3.10.4).

6.5.1.1 Adverse Events of Special Interest

Adverse events of special interest (AESIs) have been defined based on known effects of LAMAs, LABAs, and ICS. These include but are not limited to cardiovascular effects, ocular disorders, urinary retention, gastrointestinal disorders, and anticholinergic effects for LAMAs; 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, LAMAs, and LABAs are listed in Table 7. Standardized MedDRA queries will be utilized when possible and a selection of preferred terms in other situations (*Appendix 5*).

Table 7 Adverse Events of Special Interest

Medical Concept	Selection of MedDRA Terms
Adrenal suppression	Adrenal cortical hypofunction HLT
Agitation or anxiety	Collection of PTs
Anticholinergic effects	Anticholinergic syndrome SMQ Dry mouth PT
Bone fracture	Collection of HLGTs, HLTs, and PTs.
Candidiasis	Collection of PTs
Cardiovascular	Cardiac arrhythmias SMQ Cardiac failure SMQ Ischemic heart disease SMQ
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 condition	Gastrointestinal perforation, ulceration, haemorrhage or obstruction SMQ
Headache	Headaches (HLGT)
Hyperglycemia*	Hyperglycaemia/new onset diabetes mellitus (SMQ)
Hypercortisolism	Collection of PTs
Hypertension	Blood pressure ambulatory increased (PT) Blood pressure increased (PT) Blood systolic increased (PT)
Hypokalemia	Collection of PTs

Medical Concept	Selection of MedDRA Terms
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 Cataract collection of PTs
Osteoporosis and osteopenia	Osteoporosis/osteopenia (SMQ)
Palpitation	Palpitations PT
Paradoxical bronchospasm	Collection of PTs
Pneumonia	Collection of PTs
Psychiatric effects	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) Hypertrichosis (collection of PTs) Alopecia (collection of PTs) Skin infection (collection of PTs)
Sleep effects	Initial insomnia (PT) Insomnia (PT) Sleep disorder (PT)
Sudden death and cardiovascular death	Collection of PTs
Throat irritation	Collection of PTs
Tremor	Tremor (HLT)
Urinary retention	Urinary retention (PT)
Weight gain	Collection of PTs

Medical Concept	Selection of MedDRA Terms

Abbreviations: CNS=central nervous system.

* This medical concept will be called "diabetes mellitus" in the CSR post-text tabulations.

Appendix 5 (which will be based on the latest version of MedDRA available at the time of database lock) provides detail on selection of terms (narrow/wide designations for preferred terms are provided).

Adverse Events in MedDRA SMQs/Groupings of Interest by Term will be tabulated *(Table 3.2.3)*.

6.5.1.2 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 (*Table 3.13.3*) and by the following time intervals: 0 to <= 24 weeks and > 24 weeks (*Tables 3.13.4, 3.13.5,* and *3.13.6)*. The assessment of MACE events will include the rate of confirmed MACE events (*Table 3.13.2*). Adjudicated MACE events will be listed in *Listing 7.4*.

6.5.1.3 Pneumonia Events Determined by Adjudication Committees

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 by treatment group (*Table 3.13.3*) and by the following time intervals: 0 to <= 24 weeks and > 24 Weeks (*Tables 3.13.4 and 3.13.5)*. The assessment of pneumonia events will include the rate of confirmed pneumonia events (*Table 3.14.2*). 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 (*Table 3.14.3*). Specific patient risk factors (baseline FEV_1, age, medical history of pneumonia in the last 5 years [Yes or No], BMI, and baseline exacerbation history [1 vs. >=2]) will be evaluated for inclusion.

6.5.1.4 Cause of Death Determined by Adjudication Committees

Causes of death will be listed (*Table 3.15.2*) by subject and summarized by treatment for (1) all-cause mortality, (2) mortality of probable cardiovascular cause, (3) mortality of probable respiratory cause and (4) 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 the post-treatment period *(Table 3.1)*. The incidence of subjects with an on-treatment death event will be tabulated by adjudicated CRF category and treatment (*Table 3.15.1.1*). A similar tabulation will also be done for the post-treatment period (*Table 3.15.1.2*). 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 (*Table 3.15.3*). Adjudicated death events will be listed in *Listing 7.4*.

6.5.2 Clinical Laboratory Measurements

Lab parameters collected include the following:

Table 8 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	Albumin
Aspartate aminotransferase	Calcium[a]
Alkaline phosphatase	Chloride[a]
Bilirubin, total	Cholesterol
Gamma-glutamyl transferase	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 14 or Treatment Discontinuation	
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. A laboratory abnormality that leads to a dose-limiting toxicity (e.g., an abnormality that results in study drug dose reduction, suspension or discontinuation)
b. A laboratory abnormality that results in any therapeutic intervention (i.e., concomitant medication or therapy)
c. 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, LabCorp Laboratories (*Appendix 4*).

The baseline measurement for a laboratory parameter will be the last available measurement prior to the start of dosing at Day 1 (Visit 4).

Table 9 Analysis Study Time Window for Clinical Lab Assessments

Calculated Study Time Window	Time Interval for the Study Time Window
Pre-dose 1 hr.	$\geq$ 0 min. prior to dose
Post-dose 2 hrs.	> 0 min. to < 4 hrs. post-dose

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

The laboratory-value windows will be applied only for calcium, chloride, glucose, potassium, and sodium (i.e. the laboratory parameters that are sometimes assessed post-dose) and these windows will be applied only at the following visits: Visit 4 (Day 1), Visit 5 (Week 4), Visit 10 (Week 24), Visit 14 (Week 52), and the Treatment Discontinuation/Withdrawal Visit. For other laboratory parameters and other visits, windows will not be applied. The rationale is that for other laboratory parameters and other visits, post-dose assessments are not to be made. If there are multiple laboratory values for the same parameter at pre-dose of a visit or within the same post-dose study time window (if applicable) at 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-visit values and unscheduled-visit values are candidates for the end-of-treatment summary. 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 pH and specific gravity, respectively). Data from unscheduled visits will not be used for the by-visit summaries but both scheduled-visit data and unscheduled-visit data are candidates for clinically significant values, for the end-of-treatment summary, and for shift tables. Shift tables will be produced using the categories defined by the Common Terminology Criteria for Adverse Events (CTCAE) Version 4.03 grades (*Tables 3.16.5, 3.16.6,* and *3.16.7* for hematology, chemistry, and kidney function, 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 10 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 (where eGFR denotes estimated glomerular filtration rate)	Decrease from baseline in ≥ 1 CTCAE grade and ≥ 20% change from baseline (#)
AST(aspartate aminotransferase)	>3 x ULN
ALT (alanine aminotransferase)	>3 x ULN
Sodium	< 120 mmol/L
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 no history of diabetes, > 27.8 mmol/L (>500 mg/dL) regardless of baseline
Serum Potassium	<3.0 mmol/L
	>6.0 mmol/L

*CTCAE 5.0 criteria are based on fasting glucose values. However, subjects were not required to fast prior to obtaining blood glucose values.
(#) Only a 20% decrease from baseline is considered here: a 20% increase would not be of concern.

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 separately (*Tables 3.16.9* and *3.16.10*). For all laboratory parameters other than glucose and potassium noted in Table 8, all laboratory data for the parameter identified as potentially clinically significant for a subject will be listed (*Table 3.16.11*).

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 mean of all available pre-dose measurements taken prior to the start of dosing at the Randomization Visit (Visit 4). 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 11 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

Potentially clinically significant changes in heart rate will be assessed as follows:

Table 12 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 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, 14.2, 1.4.3* and *1.4.4* for the mITT, PP, RVU, and Safety Populations, respectively). The ITT Population does not need to be summarized because it is the same as the mITT Population at baseline. The Safety Population may not be needed either if it is the same as the mITT Population.

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. Baseline will be defined as the mean of the values prior to dosing at Visit 4 (Day 1). These summaries (*Table 3.17.1*) 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 hour, and post-dose 30 minutes and 2 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-visit data and unscheduled-visit data are candidates 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 13. No hypothesis tests will be performed.

Table 13 Analysis Study Time Windows for Vital Signs Assessments

Calculated Study Time Window	Time Interval for the Study Time Window
Pre-dose	≥ 0 min. prior to dose
Post-dose 30 min *	> 0 to <75 min. post-dose
Post-dose 2 hrs. *	≥ 75 min. to < 4 hrs. post-dose
Post-dose 12 hrs.	≥ 8 hrs. to < 16 hrs. post-dose

Note that minutes are rounded to the nearest whole number before applying time windows.
* The post-dose assessments are for the PFT sub-study only. If there are multiple vital sign values for the same parameter at pre-dose assessments after Visit 4 or within the same post-dose study time window (if applicable) at a visit, the last value will be chosen for analysis.

Data from unscheduled visits will not be used for the by-visit summaries but both scheduled-visit data and unscheduled-visit data are candidates for clinically significant values and for the end-of-treatment summary.

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 11 and Table 12 (*Table 3.17.2*).

All vital sign assessments for subjects with potentially clinically significant values will be listed (*Tables 3.17.3* and *3.17.4*).

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 mean of the pre-dose measurements taken prior to the start of treatment at the randomization visit (Visit 4). 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, where baseline is defined as the mean of the pre-dose measurements taken prior to the start of treatment at Visit 4 (Day 1). These assessments will be tabulated for each treatment and each scheduled nominal time point (derived using the time intervals for the study time windows detailed below in Table 14) at each visit and at end of treatment (*Table 3.18.1*). "End of Treatment" will be summarized for each scheduled post-baseline time point (pre-dose 1 hour, post-dose 30 minutes, and post-dose 2 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-visit data and unscheduled-visit data are candidates 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 (*Figure 3.18.1a* and *Figure 3.18.1e*). ECG data from subjects with pacemakers will not be included in analyses, but will be listed.

Data from unscheduled visits will not be used for the by-visit summaries but both scheduled-visit data and unscheduled-visit data are candidates for clinically significant values and for the end-of-treatment summary.

Table 14 Analysis Study Time Window for ECG Assessments

Calculated Study Time Window	Time Interval for the Study Time Window
Pre-dose 1 hr.	≥ 0 min. prior to dose
Post-dose 30 min. *	> 0 to < 75 min. post-dose
Post-dose 2 hrs. *	≥ 75 min. to < 4 hrs. post-dose

Note: The minutes are rounded to the nearest whole number before applying time windows.
* The post-dose assessments are for the PFT sub-study only.

If there are multiple ECG values for the same parameter at pre-dose of a visit date (other than for Visit 4) or within the same post-dose study time window on a visit date, the last value will be chosen for analysis.

Table 15 Criteria for PCS ECG Values

Parameter	Post-Baseline Criteria
QTcF Prolongation	(1) ≥ 500 msec if < 500 msec at study baseline and ≥ 30 msec change from study baseline
	(2) ≥ 530 msec if ≥ 500 msec at study baseline and ≥ 30 msec change from study baseline
	(3) Change of ≥ 60 msec from study baseline regardless of initial value

msec denotes milliseconds.

Potentially clinically significant ECG parameter values will be identified based on criteria listed in Table 15. 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*). 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
- 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 healthcare 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 healthcare 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 24-Hour Holter Monitoring

Holter monitoring will be conducted over 24 hours in a subset of approximately 800 randomized subjects (200 subjects from each treatment arm) at Visit 3 (Holter Baseline) and Visit 8 (Week 16). If there is not at least 18 hours of acceptable quality monitoring for a given assessment, then the assessment is to be repeated. In these cases, the second assessment will be used whether for baseline and/or for the Visit 8 (Week 16) value. However, any incidence of AEs indicated by the incomplete Holter findings at Week 4 will be captured. Participation in the Holter Monitoring Sub-study is listed in *Listing 9.7*.

Note that iCardiac will provide mean 24-hour heart rate, mean daytime heart rate, and mean nighttime heart rate, which will be an average across hourly estimates collected during the specific Holter monitoring period (24-hour, daytime, nighttime).

6.5.6.1 24-Hour Holter Monitoring Data Analysis

6.5.6.2 Primary Analysis

The change from baseline in mean 24-hour heart rate (HR) obtained using Holter monitoring at Week 16 will be analyzed using an ANCOVA model to evaluate treatment differences with baseline mean 24-hour HR (obtained during 24-hour Holter monitoring at screening) as a covariate. LS means and estimated treatment differences with 95% CIs will be provided (*Table 3.19.1* and *Listing 9.3.1*). The raw mean values and change from baseline values will also be summarized by treatment. A schematic box-plot will display the distribution of change from

baseline in mean heart rate by treatment, with extreme values that are 1.5*IQR above/below the upper/lower quartiles identifiable, where IQR is the interquartile range (*Figure 3.19.1*).

6.5.6.3 Secondary and Other Holter Monitoring Data Safety Analyses

The changes from baseline at Week 16 (Visit 8) for the mean daytime (06:00 to 22:00) HR, mean nighttime (22:00 to 06:00) HR, maximum 24-hour HR, and minimum 24-hour HR will be summarized and analyzed in a similar manner to the primary Holter endpoint (*Tables* and *Figures 3.19.2, 3.19.3, 3.19.4.1,* and *3.19.5.1* and *Listing 9.3.1*).

A frequency distribution of the following will be provided:

- Proportion of subjects with maximum heart rate during treatment of > 180, > 160 - ≤ 180, > 140 - ≤ 160, > 120 - ≤ 140, > 100 - ≤ 120, and 100 or less (*Table 3.19.4.2*).
- The number and percentage of subjects who had sustained ventricular tachycardia (defined as PVCs lasting > 30 seconds) will be tabulated by treatment (*Table 3.19.9*).
- Proportion of subjects with minimum heart rate during treatment of > 60, > 50 - ≤ 60, > 40 - ≤ 50, and ≤ 40 (*Table 3.19.5.2*)

The change from baseline in the number of Holter events will be summarized descriptively (mean, median, range, etc.). This analysis will be performed for change from baseline for the following parameters (calculated per hour): number of isolated ventricular events (PVCs), number of ventricular couplets, number of ventricular runs, number of isolated supraventricular events, number of supraventricular couplets, and number of supraventricular runs (*Tables* and *Figures 3.19.6.1, 3.19.7, 3.19.8, 3.19.10, 3.19.11,* and *3.19.12,* respectively, *Listing 9.3.2* for ventricular and supraventricular events).

The ventricular variables will each be analyzed using an ANCOVA to evaluate treatment differences with baseline value (from the acceptable 24-hour Holter assessment at Visit 3 (Screening) of the respective variable as a covariate. The supraventricular variables will each be analyzed using an ANCOVA to evaluate treatment differences with baseline value of the respective variable as a covariate. LS means and estimated treatment differences with 95% CIs will be provided. The raw mean values and change from baseline values will also be summarized by treatment. Note that for these analyses, the distribution of the frequency of these events for baseline (Visit 3) and at Week 16 will be evaluated for each event. If necessary, the Wilcoxon Rank Sum Test will be performed to compare treatments pairwise; a Hodges-Lehmann 95% CI for the location shift (median treatment difference) will be provided.

Additionally, the following will be tabulated:

- The proportion of subjects in each category of change from baseline in the number of PVCs per hour (no change, increase of > 0 - < 60, ≥ 60 - < 120, and ≥ 120, and decrease of > 0 - < 60, ≥ 60 - < 120, and ≥ 120) (*Table 3.19.6.2*).

- The number and percentage of subjects who had sustained ventricular tachycardia (defined as PVCs lasting > 30 seconds) will be tabulated by treatment (*Table 3.19.9*).
- The number of subjects who had (atrial fibrillation or atrial flutter) with rapid ventricular response will be tabulated by treatment (*Table 3.19.13*).
- Bradycardia and tachycardia episodes and Holter assessment will be listed in *Listing 9.3.3.* Holter monitoring withdrawal criteria met will be listed in *Listing 9.3.4.*

6.5.7 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.

6.5.8 Subgroup Analyses for Safety

Selected tabulations (named below) will be provided for the following subgroups:

- China
- Asia (Asia is defined by country rather than by race and includes China, Japan, South Korea and Taiwan)
- US
- Europe (including Russia)
- North America
- Other
- Baseline Blood Eosinophil Count (< 150 cells per mm^3 or ≥ 150 cells per mm^3) (Visit 1)
- Racial groups:
  - Black or African American
  - White
  - Asian
  - Other, Native Hawaiian or Other Pacific Islander, or American Indian or Alaska Native
- Age groups:
  - Age < 65 years
  - Age >= 65 years
  - Age >= 65 – <75 years
  - Age >=75 years
- Sex groups:
  - Male

- Female

Note that more age categories are being used for safety analyses than for efficacy analyses.

Safety subgroup summaries will also be performed for AEs (*Tables 5.1.1* to *5.2.1.11f*) and SAEs (*Tables 5.3.1.1* to *5.3.1.12f*) by subgroup.

7. CHANGES FROM METHODS PLANNED IN THE PROTOCOL

Changes to the following statistical methods planned in Version 6 of the protocol are as follows:

- Removal of the following other efficacy endpoints:
  - Rate of COPD exacerbations treated with systemic steroids, Rate of COPD exacerbations treated with antibiotics, Time to first COPD exacerbation treated with systemic steroids, and Time to first COPD exacerbation treated with antibiotics.
  - EXACT Respiratory Symptoms (E-RS) symptom domain scores for breathlessness, cough and sputum, and chest symptoms over 24 weeks, over 52 weeks, and over each 4-week interval of the 52-week treatment period.
  - Individual components of the TDI: functional impairment, magnitude of task, and magnitude of effort over 24 weeks, over 52 weeks, and at each post-randomization visit.
  - Individual domains of the SGRQ: Symptoms, Activity, and Impacts over 24 weeks, over 52 weeks, and at each post-randomization visit.
  - Sub-study analyses for FVC, PEFR and FEF_{25-75}.
- The rate of severe COPD exacerbations is switched from being an other endpoint to being a secondary endpoint.

8. STATISTICAL SOFTWARE

Data processing, statistical screening, descriptive reporting and analysis of the efficacy and safety data will be performed using SAS (Version 9.4 or higher). Graphs may also be produced using R (R Development Core Team, 2003).

9. REFERENCES

Anzures-Cabrera, J. and Higgins, J. P. T. Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, 2010;1: 66–80.

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.

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).

Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. (2016), Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. Office of Health Economics, Research Paper 16/01.

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

Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease. Updated; 2014:1-102. Available from: http://www.goldcopd.org/uploads/users/files/GOLD_Report2014_Feb07.pdf

Gottlow M, Hollis S, Wan R, Hirsch I, Darilay A, Weissfeld L, France L. A Simulation study of a controlled imputation approach for analyzing missing data in recurrent events due to early discontinuations. PSI Annual Conference 2015.

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.

Jennison C, Turnbull BW. Group Sequential Methods: Applications to Clinical Trials. Chapman and Hall/CRC: Florida, 2000.

Keene ON, Roger JH, Hartley BF. Kenward MG. Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharmaceutical Statistics*, 2014;13: 258-264.

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.

Meng XL. Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 1994;9(4):538-558.

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

Moscovici JL, Ratitch B. Combining Survival Analysis Results after Multiple Imputation of Censored Event Times. PharmaSUG 2017 – Paper SP05, 2017. Available from https://www.pharmasug.org/proceedings/2017/SP/PharmaSUG-2017-SP05.pdf

Mütze, T., Glimm, E., Schmidli, H., & Friede, T. (2018). Group sequential designs for negative binomial outcomes. Statistical Methods in Medical Research. https://doi.org/10.1177/0962280218773115

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 Framework 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)

von Hippel, P.T. (2017). "Maximum likelihood multiple imputation: Faster imputations without posterior draws." arXiv:1210.0870

Wan R, Hirsch I, Gottlow M, Hollis S, Darilay A, Weissfeld L, France L. Controlled imputation approach for analyzing missing data in recurrent events due to early discontinuations. DIA/FDA Statistics Forum 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.