Statistical Analysis Plan I I8F-MC-GPIF Version 3

A Master Protocol to Investigate the Efficacy and Safety of Tirzepatide Once Weekly in Participants who have Obstructive Sleep Apnea and Obesity: A Randomized, Double-Blind, Placebo-Controlled Trial (SURMOUNT-OSA)

NCT05412004

Approval Date: 28-FEB-2024

Statistical Analysis Plan I8F-MC-GPIF: Efficacy and Safety of Tirzepatide Once Weekly in Participants who have Obstructive Sleep Apnea and Obesity: A Randomized, Double-Blind, Placebo-Controlled Trial (SURMOUNT-OSA)

Protocol Title: A Master Protocol to Investigate the Efficacy and Safety of Tirzepatide Once Weekly in Participants who have Obstructive Sleep Apnea and Obesity: A Randomized, Double-Blind, Placebo-Controlled Trial (SURMOUNT-OSA)

Protocol Number: I8F-MC-GPIF

Compound Number: Tirzepatide (LY3298176)

Short Title: A Statistical Analysis Plan for Tirzepatide in Participants with Obstructive Sleep Apnea and Obesity

Sponsor Name: Eli Lilly and Company

Legal Registered Address: Indianapolis, Indiana USA 46285

Regulatory Agency Identifier Number(s)

IND	157090

Document ID: VV-CLIN-117998

Approved on 28 Feb 2024 GMT

Table of Contents

Version History

This statistical analysis plan (SAP) is the second version and is based on amendment (c) of the protocol for I8F-MC-GPIF (GPIF) approved on 02 June 2023. This SAP was approved prior to the first unblinding of the treatment assignments for the primary outcome lock.

Table GPIF.1.1. SAP Version History Summary

SAP Version	Approval Date	Change	Rationale
1	12 January 2023	Not Applicable	Original version
2	17 October 2023	**Section 1.1:** • Revised the primary endpoint from percent change in AHI to change in AHI and added percent change in AHI to key secondaries.	• Changed to align with regulatory recommendation.
		• Moved hypoxic burden from secondary to key secondary endpoint.	• Change made due to the increasing importance of hypoxic burden in OSA disease state.
		• Moved FOSQ to secondary from key secondary endpoints and added PROMIS score related endpoints to key secondary.	• Changed to reflect regulatory recommendation.
		Section 1.1.1: • Added language to clarify the population and intercurrent events for the estimands.	• Added for clarification in alignment with regulatory feedback.
		Section 2.1: • Added detailed multiplicity control scheme for controlling Type 1 error.	• Details of Type 1 error control provided as planned in the protocol.

SAP Version	Approval Date	Change	Rationale
		Section 3: • Modified language on analysis sets to clarify the definition of the analysis sets and population.	• Analysis set definitions updated for clarity.
		Section 4.1: • Updated the definition of baseline and postbaseline measures for safety analyses. • Updated baseline and postbaseline definition for PRO measures.	• Changed to minimize missing baseline data relevant to dosing. • Updated to reflect the proper collection time of PROs associated with a PSG measurement. Added a 7-day window for each PRO visit to minimize missing data.
		Section 4.1.2: • Updated the intercurrent events in Table GPIF.4.2	• Change made to clarify the definition of intercurrent events.
		Section 4.1.6: • Removed analysis of changes to baseline medication in postrandomization (in term of type/class): o lipid lowering therapy, and o antihypertensive therapy.	• Data not collected.
		Section 4.3.2: • Added option to include interaction term between treatment and covariates in ANCOVA model.	• Change made in alignment with industry guidance for handling covariates.

SAP Version	Approval Date	Change	Rationale
		Section 4.4.1: • Updated primary endpoint.	• Changed to align with the endpoint change in Section 1.1
		Section 4.4.1: • Updated key secondary endpoints.	• Changed to align with the endpoint change in Section 1.1
		Section 4.4.1.1: • Removed hierarchical endpoint for FOSQ and added hierarchical endpoint for PROMIS score. • Added an option to impute missing baseline PRO with multiple imputation.	• Changed to align with the endpoint change in Section 1.1. • Added to mitigate the effect of missing data on PRO measurements.
		Section 4.4.1.3: • Added an option to use tipping point analysis as a sensitivity analysis.	• Changed to align general regulatory recommendations for handling missing data.
		Section 4.4.3: • Added analysis methods for secondary endpoints not controlled for Type 1 error.	• Changed to align with the endpoint change in Section 1.1.
		Section 4.6.3.3.2: • Updated analysis for hepatic safety.	• Changed to align to the standardized analysis approach across tirzepatide indications.
		Section 4.6.3.4: • Updated severe hypoglycemia definition.	• Updated to align with the protocol.

SAP Version	Approval Date	Change	Rationale
		Section 4.6.3.5.4: • Removed specific immunogenicity analyses.	• Some immunogenicity analyses are planned as an integrated summary instead of at individual study level and are removed.
		Section 6.4: • Added Appendix 4.	• Added to prespecify statistical analyses for China subpopulation.
		Section 6.5: • Added Appendix 5.	• Added to prespecify statistical analyses for Japan subpopulation.
3	See Date on Page 1	**Section 1.1:** • Moved the hierarchical combination of PROMIS endpoints from key secondary to other secondary. • Added change in FOSQ (30-item) total score in other secondary endpoints	• Changed per FDA recommendation. • Added to clarify that the overall as well as by domain analysis will be performed
		Section 1.1.1: • Identified the treatment regimen estimand as the primary estimand for marketing application. • Data from inadvertent enrollees are to be included in primary and key secondary analysis.	• Clarified per FDA suggestion. • Changed per FDA recommendation.

SAP Version	Approval Date	Change	Rationale
		• Updated the subsections titled "Handling of intercurrent events" under estimand definition.	• Clarified language per FDA feedback.
		Section 2.1: • Updated graphical testing strategy and Figure GPIF.2.1.	• Graphical testing strategy revised per FDA recommendation. PROMIS related endpoints removed from individual study graph and included in the integrated efficacy analysis subject to submission wide type 1 error rate control
		Section 3: • Changed the definitions for data point sets.	• Changed to align with changes in Section 1.1.1.
		Section 4.1: • Added language on baseline AHI. • Modified the postbaseline definition for PRO measures. • Updated that geographic region will be used in lieu of pooled country as a covariate in analysis models. • Changed AHI groups to be included in the model as covariates.	• Language added for clarification. • Updated to clarify the definition. • Changed to reduce the number of strata in the model and to keep enough participants in each covariate strata. • Changed the group category from "moderate" to "not severe" to incorporate inadvertently enrolled participants.

SAP Version	Approval Date	Change	Rationale
		Section 4.1.2: • Updated missing data and imputation method in Table GPIF.4.2.	• Missing data imputation algorithm revised, in part to include inadvertent enrollees in analysis, and language on missing data clarified per FDA feedback.
		Section 4.1.2: • Added no OSA and mild categories to baseline OSA category	• Added to account for inadvertently enrolled patients
		Section 4.1.6: • Added definition for baseline and postbaseline concomitant medication use.	• Clarified definition for related tables and listings.
		Section 4.2: • Added definition for participant study disposition.	• Clarified the definition of participant study disposition based on collected CRF data.
		Section 4.4.1.2: • Added analysis for change in log hypoxic burden. • Added reporting of unconditional risk difference from logistic regression.	• Log scale is deemed appropriate for analysis as the measure is an area under the curve. • Added to provide additional measure of treatment effect in accordance to FDA guidance.

SAP Version	Approval Date	Change	Rationale
		Section 4.6: • Added hepatobiliary events, malignancies, and abuse potential as special safety topics.	• Updated to provide information on special safety topics.
		Section 4.6.1.1: • Clarified definition for TEAEs.	• Clarification
		Section 4.6.6: • Added information on product complaints	• Added in accordance with regulatory requirements for devices or combination products
		Section 4.7.2.4: • Updated method of calculation for PROMIS T-scores	• Updated to calculate T-scores using a response pattern scoring approach in accordance with FDA feedback.
		Section 4.7.2.5: • Updated method of calculation for PROMIS T-scores	• Updated to calculate T-scores using a response pattern scoring approach in accordance with FDA feedback.
		Section 5: • Added sample size determination from protocol	• Added in accordance with FDA feedback
		Section 6.3: • Added language that no imputation will be performed for MRI data.	• Clarification

SAP Version	Approval Date	Change	Rationale
		Throughout the document • Changed GPI1 to ISA1 and GPI2 to ISA2	• Changed for consistency

Abbreviations and Definitions

Term	Definition
ADA	anti-drug antibody
AE	adverse event
AESI	adverse event of special interest
AHI	Apnea-Hypopnea Index
ANCOVA	analysis of covariance
BG	blood glucose
BMI	body mass index
CRF	case report form
CSR	clinical study report
C-SSRS	Columbia-Suicide Severity Rating Scale
EAS	efficacy analysis set
eCRF	electronic case report form
EQ-5D-5L	EuroQol-5 Dimension-5 Level
ESS	Epworth Sleepiness Scale
EudraCT	European Union Drug Regulating Authorities Clinical Trials
FAS	Full Analysis Set
FDA	United States Food and Drug Administration
FOSQ	Functional Outcomes of Sleep Questionnaire
FOSQ-10	Functional Outcomes of Sleep Questionnaire, 10 items
GI	gastrointestinal
GIP	glucose-dependent insulinotropic polypeptide
GIPR	glucose-dependent insulinotropic polypeptide receptor
GLP-1	glucagon-like peptide-1
GLP-1R	glucagon-like peptide-1 receptor
HLT	High Level Term

Approved on 28 Feb 2024 GMT

Term	Definition
ISA	intervention-specific appendix
ISR	Injection site reaction
JASSO	Japan Society for the Study of Obesity
Lilly	Eli Lilly and Company
LLT	Lowest Level Term
MACE	major adverse cardiovascular event(s)
MedDRA	Medical Dictionary for Regulatory Activities
mITT	modified intent-to-treat
MMRM	mixed model repeated measures
MRD	minimum required dilution
MRI	magnetic resonance imaging
MTD	maximum tolerated dose
Nab	neutralizing antibodies
Nab-	neutralizing antibody negative
Nab+	neutralizing antibody positive
Nab LY	tirzepatide
OSA	obstructive sleep apnea
OSAS	obstructive sleep apnea syndrome
OUS	outside of the United States
PAP	positive airway pressure
PGIC	Patient Global Impression of Change
PGIS	Patient Global Impression of Status
PHQ-9	Patient Health Questionnaire
PK	pharmacokinetic
PRO	patient-reported outcome
PROMIS	Patient-Reported Outcomes Measurement Information System

Term	Definition
PSG	polysomnography
PT	Preferred Term
REML	restricted maximum likelihood
SAE	serious adverse event
SAP	statistical analysis plan
SBP	systolic blood pressure
SF-36v2	Short-Form 36 version 2
SMQ	Standardized MedDRA Query
SOC	System Organ Class
TE ADA	treatment-emergent anti-drug antibody
TE ADA-	treatment-emergent anti-drug antibody negative
TE ADA+	treatment-emergent anti-drug antibody positive
TEAE	treatment-emergent adverse event
UACR	urine albumin-to-creatinine ratio
ULN	upper limit of normal

1. Introduction

1.1. Objectives, Endpoints, and Estimands

Objective	Endpoints
Primary	
To demonstrate that tirzepatide at the MTD (10 mg or 15 mg) QW is superior to placebo for decrease in AHI.	Change in AHI from baseline to Week 52.
Key Secondary (controlled for Type 1 error)	
To demonstrate that tirzepatide at the MTD (10 mg or 15 mg) QW is superior to placebo for	From baseline to Week 52
Percent change in AHI	Percent change in AHI
Clinically meaningful change in AHI	Percent of participants with ≥50% AHI reduction
Achieving OSA remission or mild non-symptomatic OSA	Percent of participants with AHI <5 or AHI 5-14 with ESS ≤10
Change in body weight	Percent change in body weight
Change in inflammatory status	Change in hsCRP concentration
Hypoxic burden	Change in SASHB (% min/hour)
Change in PROs	Change in [a]: PROMIS Sleep-related impairment short form 8a PROMIS Sleep disturbance short form 8b
Change in SBP	From baseline to Week 48[b] Change in SBP
Other Secondary	
To demonstrate that tirzepatide at the MTD (10 mg or 15 mg) QW is superior to placebo for	From baseline to Week 52
Change in excessive daytime sleepiness	Change in ESS score
Change in patient-reported functional status as assessed by FOSQ (30 items)	Change in FOSQ-10 score Change in FOSQ (30 items) Score Change in all FOSQ domain scores, specifically General Productivity Activity level Vigilance Social outcomes Intimate and sexual relationships
Change in body weight	Percent of participants who achieve ≥10% body weight reduction ≥15% body weight reduction ≥20% body weight reduction
Change in lipid parameters	Change in HDL-cholesterol non-HDL-cholesterol triglycerides

Objective	Endpoints
A hierarchical assessment of PRO change	A hierarchical combination of the following: Change in PROMIS Sleep-related impairment short form 8a Change in PROMIS Sleep disturbance short form 8b
Change in supportive secondary PROs	Change in: SF-36v2 acute form domain and summary scores Percent of participants with improved categorical shift in: PGIS-OSA Sleepiness PGIS-OSA Fatigue PGIS-OSA Snoring Proportion of participants achieving clinically meaningful within-patient change in: PROMIS Sleep-related impairment PROMIS Sleep disturbance
Insulin Change in DBP	Change in fasting insulin From baseline to Week 48[a] Change in DBP
Exploratory	
To demonstrate that tirzepatide at the MTD (10 mg or 15 mg) QW is superior to placebo for Change in exploratory PROs	From baseline to Week 52 Change in EQ-5D-5L utility index EQ-VAS scores Percent of participants with improved categorical shift in: PGIC-OSA Sleepiness PGIC-OSA Fatigue PGIC-OSA Sleep quality PGIC-OSA Snoring
To evaluate the effect of tirzepatide on sleep parameters as measured by Actigraphy (AX6)	Change from baseline to endpoint assessment in Daytime sleep duration Daily step counts Average acceleration

Abbreviations: AHI = Apnea-Hypopnea Index; AX6 = Axivity 6; BP = blood pressure; DBP = diastolic blood pressure; ESS = Epworth Sleepiness Scale; EQ-5D-5L = EuroQol-5 Dimension-5-Level; EQ-VAS = EuroQol Visual Analogue Scale; FOSQ = Functional Outcomes of Sleep Questionnaire; HDL = high-density lipoprotein; hsCRP = high-sensitivity C-reactive protein; MTD = maximum tolerated dose; OSA = obstructive sleep apnea; PAP = positive airway pressure; PGIC-OSA = Patient Global Impression of Change – Obstructive Sleep Apnea; PGIS-OSA = Patient Global Impression of Status – Obstructive Sleep Apnea; PRO = patient-reported outcome; PROMIS = Patient-Reported Outcomes Measurement Information System; QW = once weekly; SASHB = sleep apnea-specific hypoxic burden; SBP = systolic blood pressure; SF-36v2 = Short-Form 36 version 2.

[a] Subject to submission wide type 1 error rate control (Vandemeulebroecke et al. 2024).

[b] BP will be assessed at Week 48 because PAP withdrawal at Week 52 may confound BP assessment.

1.1.1. Estimands

Primary estimands

The primary and each key secondary efficacy analysis will be guided by the "*treatment regimen*" estimand and the "*efficacy*" estimand to support global regulatory submissions and publications. The "*efficacy*" estimand provides an on-treatment assessment of efficacy without confounding the treatment effect from the data collected after treatment discontinuation. It represents on-treatment efficacy. The "*treatment regimen*" estimand estimates the treatment effect, including the effect of intervention discontinuation to reflect clinical practice. It represents the efficacy irrespective of adherence to study intervention. The "*treatment regimen*" estimand will be used as the primary estimand to support a marketing application for the FDA.

Efficacy estimand

The clinical question of interest for the efficacy estimand is the treatment difference between tirzepatide and placebo after 52 weeks of intervention in treated participants with obesity and OSA, prior to study intervention discontinuation for any reason.

Efficacy estimand attributes

- *Population:* Adult participants with obesity and OSA who received at least 1 dose of study treatment.
- *Treatment condition:* On randomized treatment.
- *Endpoints:* The primary and key secondary endpoints will be studied. Further details on the endpoints can be found in the Objectives and Endpoints table (Section 1.1).
- *Population level summary:* The difference in mean change from baseline to 52 weeks will be used for continuous endpoints; the difference in proportion (absolute or relative, as appropriate) will be used for dichotomous endpoints. The population level summary will be conducted using the EAS described in Section 3.
- *Handling of intercurrent events:* The intercurrent events of treatment discontinuation and use of PAP therapy for participants in ISA1 is addressed by the hypothetical strategy. The potential outcome of interest is the response in the efficacy measurement if participants would remain on their randomly assigned treatment for 52 weeks and would not initiate PAP therapy during the study.
- *Rationale:* The efficacy estimand provides an on-treatment assessment without confounding the treatment effect from off-treatment data.

Treatment regimen estimand

The clinical question of interest for the treatment regimen estimand is the treatment difference between tirzepatide and placebo after 52 weeks of intervention in treated participants with obesity and OSA, regardless of intervention discontinuation for any reason.

Treatment regimen estimand attributes

- *Population:* Adult participants with obesity and OSA who received at least 1 dose of study treatment.
- *Treatment condition:* On- or off-randomized-treatment.
- *Endpoints:* The primary and key secondary endpoints will be studied. Further details on the endpoints are in the Objectives and Endpoints table (Section 1.1).
- *Population level summary:* The difference in mean change from baseline to 52 weeks will be used for continuous endpoints and the difference in proportion (absolute or relative, as appropriate) will be used for dichotomous endpoints. The population level summary will be conducted using the FAS described in Section 3.
- *Handling of intercurrent events:* No intercurrent events since treatment adherence and the initiation of PAP therapy are part of the treatment condition. Methods to handle missing data are described in detail in Section 4.1.2.
- *Rationale:* The treatment regimen estimand estimates treatment effect, including the effect of intervention discontinuation to reflect clinical practice. It is used for submission and registration purpose with regulatory agencies.

Efficacy and treatment regimen estimands will be evaluated for key secondary objectives similarly to the primary objectives.

Safety estimand

The clinical interest for safety estimands is the safety assessment of individual treatment arms up to the end of safety follow-up or study discontinuation in participants with obesity and OSA, from all randomly assigned participants who are exposed to at least 1 dose of study intervention, regardless of adherence to study intervention.

Safety estimand attributes

- *Population:* Adult participants with obesity and OSA who received at least 1 dose of study treatment.
- *Treatment condition:* On- or off-randomized-treatment.
- *Endpoints:* Endpoints corresponding to the safety analyses described in Section 4.6.
- *Population level summary:* Population level summaries will be conducted using the safety analysis set described in Section 3.
- *Intercurrent events*: Potential intercurrent events may lead to study discontinuation or missing data due to a technical or scheduling issue, but there are no planned approaches for accommodating intercurrent events.

1.2. Study Design

Study I8F-MC-GPIF (GPIF) is a multicenter, randomized, parallel-arm, double-blind, placebo-controlled Phase 3 study to evaluate the efficacy and safety of tirzepatide at the MTD (10 mg or 15 mg) once weekly versus placebo in participants who have obesity and moderate to severe OSA.

This basket-type master protocol will investigate 2 participant populations, described in 2 ISAs:

- ISA1 will include participants who are unwilling or are unable to use PAP therapy.
- ISA2 will include participants who have been on PAP therapy for at least 3 consecutive months prior to Visit 1 and plan to continue PAP therapy during the study.

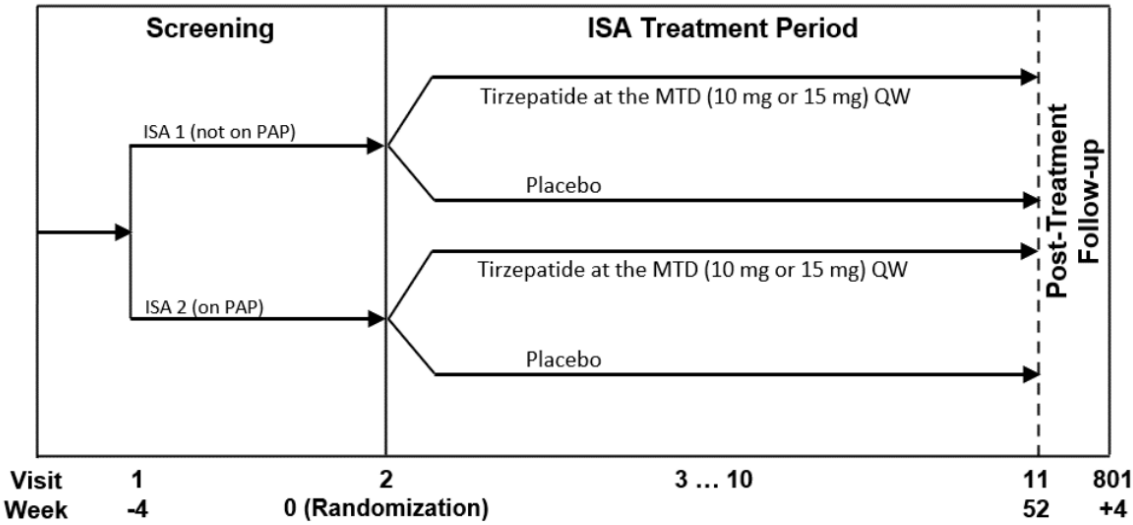
Participants to be assigned to whichever ISA they qualify for. Participants will then be randomly assigned to:

- tirzepatide at the MTD (10 mg or 15 mg) subcutaneous once weekly, or
- placebo.

The expected total duration of study participation for each participant, including screening and the posttreatment follow-up periods, is 60 weeks across the following study periods:

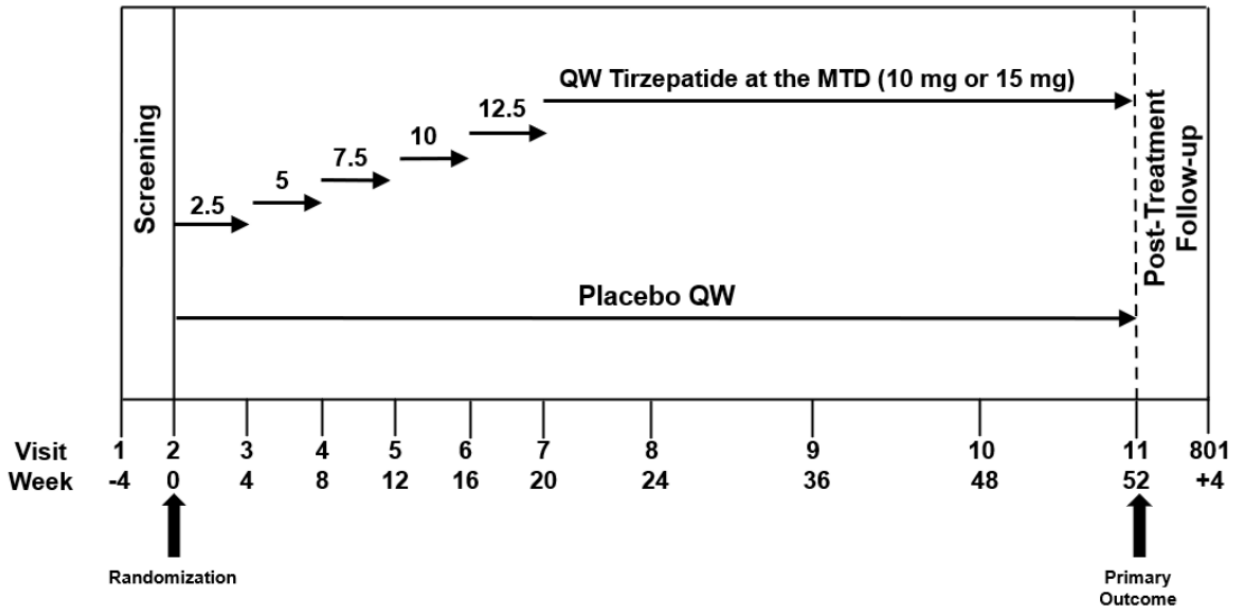
- Screening: 4 weeks
- ISA Treatment Period: 52 weeks
- Post-Treatment Follow-up Period: 4 weeks

The maximum duration of treatment is 52 weeks.



Abbreviations: ISA = intervention-specific appendix; MTD = maximum tolerated dose; PAP = positive airway pressure; QW = once weekly.

Figure GPIF.1.1. Illustration of master protocol design for Clinical Protocol I8F-MC-GPIF.



Abbreviations: MTD = maximum tolerated dose; QW = weekly.

Figure GPIF.1.2. Illustration of dose escalation and visit schema for Clinical Protocol I8F-MC-GPIF.

2. Statistical Hypotheses

For each ISA, the primary objective is to demonstrate that tirzepatide at the MTD (10 mg or 15 mg) is superior to placebo in treating participants with OSA with respect to the change in AHI. Thus, the null and alternative hypotheses will be defined as below.

Null hypothesis: tirzepatide at the MTD (10 mg or 15 mg) is not different from the placebo with respect to the mean change from baseline in AHI at 52 weeks.

Alternative hypothesis: tirzepatide at the MTD (10 mg or 15 mg) is superior to the placebo with respect to the mean change from baseline in AHI at 52 weeks.

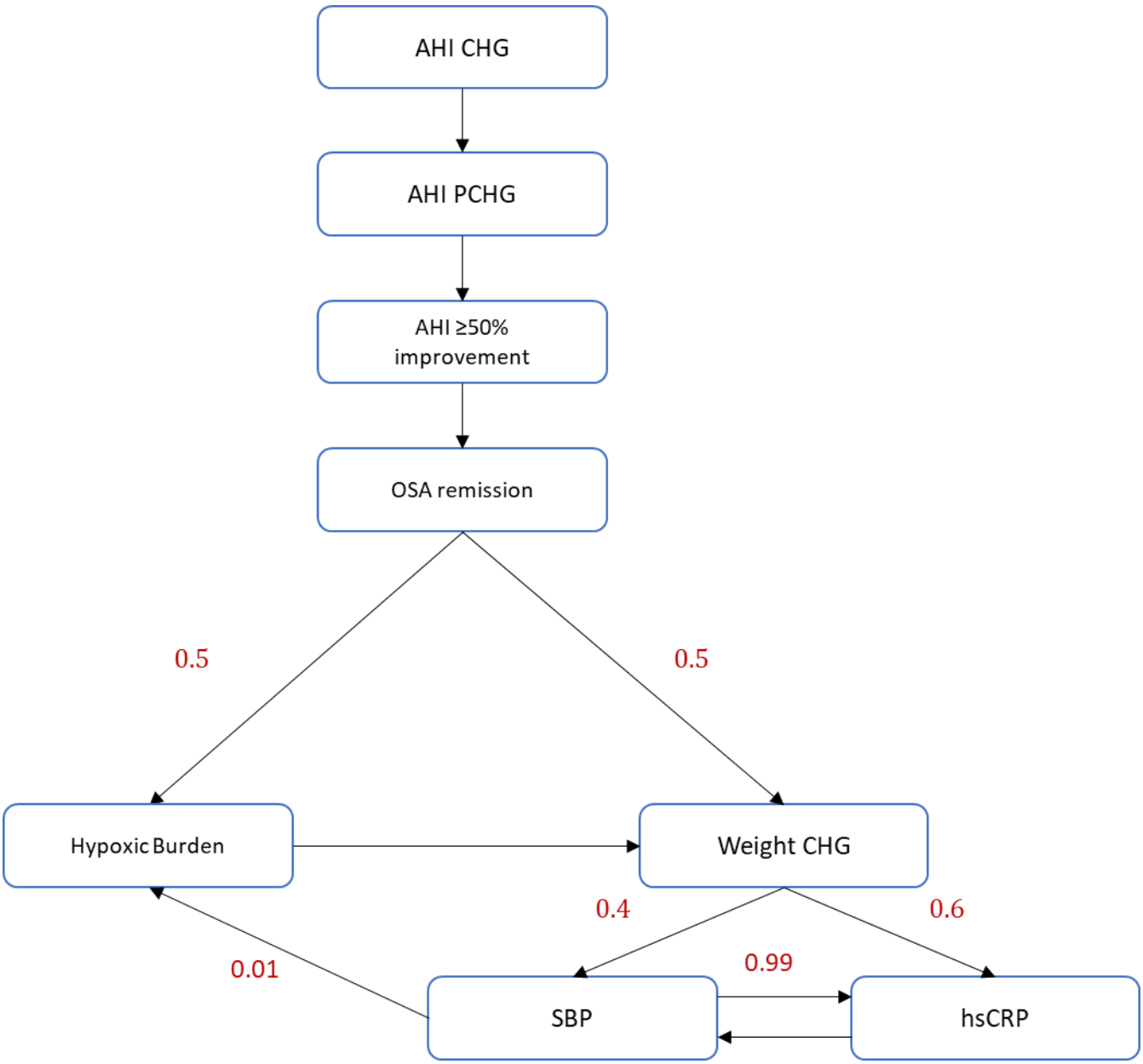
The treatment effect will be defined as the difference between the estimates of the mean change from baseline at 52 weeks for tirzepatide at the MTD (10 mg or 15 mg) and placebo. Operationally, the hypotheses will be evaluated by 2-sided tests.

2.1. Multiplicity Adjustment

Multiplicity adjusted analyses will be performed on the primary and key secondary objectives to control the overall family-wise Type 1 error rate at a 2-sided alpha level of 0.05 within each ISA. The graphical multiple testing procedure described in Bretz et al. (2009, 2011) will be used. This approach is a closed testing procedure; hence, it strongly controls the family-wise error rate across all hypotheses (Alosh et al. 2014).

Figure GPIF.2.1 provides the details of the graphical multiple testing procedure. Because the 2 types of estimands (treatment regimen and efficacy estimands) are intended for distinct purposes, no multiplicity adjustment will be made for conducting separate analyses on the same objectives. Unless otherwise specified, there will be no multiplicity adjustments for evaluating exploratory objectives and safety assessments.

Analysis for change in PROMIS Sleep-related impairment short form 8a and PROMIS Sleep disturbance short form 8b is specified in the integrated efficacy analysis plan to be tested subject to the submission wise error rate control strategy (Bretz and Xi 2019, Vandemeulebroecke et al. 2024) by conducting a pooled analysis across the 2 ISAs.



Abbreviations: AHI = Apnea-Hypopnea Index; CFB = change from baseline; CHG = change; GPIF = I8F-MC-GPIF; hsCRP = high-sensitivity C-reactive protein; OSA = Obstructive Sleep Apnea; PCHG = percentage change; SBP = systolic blood pressure.

Figure GPIF.2.1 Graphical testing scheme for Study GPIF.

3. Analysis Sets

Table GPIF.3.1 describes the populations that will be used for statistical analyses within each ISA of the master protocol. Additional intervention-specific populations for analyses may be described in the respective ISA.

Table GPIF.3.1. Description of Analysis Population

Analysis Population	Description
Entered	All participants who sign informed consent.
Randomized	All participants who are randomly assigned a study treatment (double-blind).
Modified intent-to-treat (mITT)	All randomized participants who are exposed to at least 1 dose of study intervention.

Table GPIF.3.2. Description of Analysis Data Point Sets

Analysis Set	Description
Full analysis set (FAS)	Data obtained during treatment period of set of participants from the mITT population, regardless of adherence to study intervention.
Efficacy analysis set (EAS)	Data obtained during treatment period of set of participants from the mITT population, excluding data after discontinuation of study intervention (last dose + 7 days) and for ISA1, excluding data after initiating PAP therapy.
Safety analysis set (SS)	Data obtained during treatment and safety follow-up period of set of participants from the mITT population, regardless of adherence to study intervention.

Abbreviations: EAS = efficacy analysis set; FAS = full analysis set; ISA = intervention-specific appendix; mITT = modified intent-to-treat; PAP = positive airway pressure; SS = safety analysis set.

4. Statistical Analyses

4.1. General Considerations

Statistical analysis will be the responsibility of Eli Lilly and Company (Lilly) or its designee. Statistical analysis for each ISA will be conducted individually and a combined analysis with both ISAs is not planned. All analyses specified will apply to both ISAs unless the analysis is specified as ISA-specific.

The SAP will be finalized prior to the unblinding of the first ISA.

Some analyses and summaries described in this analysis plan may not be conducted if not warranted by data (for example, too few events to justify conducting an analysis). Additional exploratory analyses of the data will be conducted as deemed appropriate.

Efficacy analyses will be conducted on all participants randomly assigned to study intervention according to the treatment to which the participants are assigned and were exposed to at least 1 dose. For the "treatment regimen" estimand, the analysis will be conducted using the FAS. To minimize missing data, participants randomly assigned to study intervention who prematurely discontinue study treatment will be encouraged to remain in the study. However, some participants may choose to permanently discontinue from the study which will lead to missing endpoints. Details on handling missing values can be found in Section 4.1.2. For the "efficacy" estimand, the analysis will be conducted using the EAS.

Safety analysis will be conducted using the Safety Analysis Set. Selected safety analyses may be conducted after excluding the data after permanent discontinuation of the study intervention. For the safety related parameters, the definition of baseline and postbaseline are specified in Table GPIF.4.1.

Table GPIF.4.1. Baseline and Postbaseline Definition for Safety Analyses

Analysis Set	Analysis Type	Baseline	Postbaseline
SS	1.1) Treatment-Emergent Adverse Events	The baseline period is defined as the start of screening and ends prior to the first dose of study treatment (typically at Week 0). If the first dose date is missing, then the randomization date will be used instead of first dose date.	Starts at or after the first dose of study treatment and ends at the end of the study period (including off-drug follow up visit).
SS	1.2) Treatment-Emergent Abnormal Laboratory Results[a] and Vital Signs	For laboratory results, baseline period is defined as prior to the first dose time and will include all scheduled and unscheduled measurements. If the first dose time is missing, then any data collected on the date of the first dose will be treated as baseline. For vital signs, baseline period is defined as measurements collected prior to the first dose.	Postbaseline will be defined as after the baseline period through the end of the study participation. All scheduled and unscheduled measurements will be included.

Analysis Set	Analysis Type	Baseline	Postbaseline
		If the first dose date is missing, then the randomization date will be used instead of first dose date.	
SS	1.3) Change from Baseline for Laboratory Results[a], and Vital Signs	The last scheduled and unscheduled nonmissing assessment recorded during the baseline period defined above (1.2).	Postbaseline will be defined as above (1.2). Only scheduled visits will be included. The ED visits are considered scheduled visits.

Abbreviations: ED = early discontinuation; SS = Safety Analysis Set.

[a] Immunogenicity related analysis is specified in Section 4.6.3.5.

For AHI analyses, baseline is defined as the last nonmissing measurement prior to the first dose.

The following paragraphs define selection of the PRO response which will be used for analysis at baseline and postbaseline visits. To select the baseline observation for PROs which are planned to be completed on the same day as the PSG (ESS, FOSQ, PROMIS, PGIS, SF-36v2 acute form, and EQ-5D-5L), the observation completed on the day or on the next day of the start of the baseline PSG will be selected. If multiple responses are completed within this period, the last response given within this timeframe will be selected. If no response was provided within this timeframe, the latest observation completed prior to the first dose will be selected. If a baseline still cannot be identified, the earliest observation within a 7-day window from the start of treatment date will be selected.

For postbaseline visits with a planned PSG measurement, the response for the PROs which are planned to be completed on the same day as the PSG (ESS, FOSQ, PROMIS, PGIS, PGIC, SF-36v2 acute form, and EQ-5D-5L) will be selected in the following way. If available, the observation completed on the day or on the next day of the start of the PSG will be selected. If multiple responses are completed within this period, then the last response given within this timeframe will be selected. If no response was provided in this timeframe, the latest observation completed within the visit window will be selected. If a measurement for the visit still cannot be identified, then the observation within a 7-day window around the start of visit date that is closest to the visit start date will be selected.

For postbaseline visits without a planned PSG measurement, the response for the PROs (ESS, FOSQ, PROMIS, PGIS, PGIC, SF-36v2 acute form, and EQ-5D-5L) will be identified by selecting the latest observation completed within the visit window. If a measurement for the visit cannot be identified in this manner, then the observation within a 7-day window around the start of the visit date that is closest to the visit start date will be selected.

To select the baseline observation for the PHQ-9 and C-SSRS, if multiple responses are completed prior to the first dose and there are no differences in these responses, the observation completed most recently, prior to the first dose, will be selected. If multiple responses are completed prior to the first dose and there are differences in these responses, the approach differs based on the questionnaire. For the PHQ-9, the response with the worst total score will be

selected; for the C-SSRS, the worst response for each question will be selected and each of the worst responses will be combined into a single response which will be used for analysis. For each postbaseline visit, the same approach to select a response in the case of multiple responses within the same visit window will be carried out.

For all other analyses, baseline is defined as the last nonmissing measurement prior to the first dose unless otherwise specified.

For AHI, if there are multiple observations for the same visit, then the later observation will be selected.

Statistical treatment comparisons will be performed between tirzepatide MTD and placebo. Unless otherwise noted, all tests of treatment effects will be conducted at a 2-sided alpha level of 0.05, and the confidence interval will be calculated at a 2-sided 95% level. In statistical summaries and analyses, participants will be analyzed as randomized. Analysis models will use geographic region (US/OUS) as a covariate when applicable.

Analysis of covariance will be used to analyze continuous variables collected only at baseline and endpoint. Unless otherwise specified, the model will include treatment and strata (geographic region [US/OUS], AHI stratum [not severe (AHI <30), severe (AHI ≥30)], and gender) as fixed effects and baseline as a covariate.

MMRMs will be used to analyze continuous variables collected at baseline and more than 1 postbaseline visit. For the MMRM analysis, REML will be used to obtain model parameter estimates for continuous longitudinal variables. All the longitudinal observations at each scheduled postbaseline visit will be included in the analysis. The model will include the fixed class effects of treatment, strata (geographic region [US/OUS] and gender), visit, and treatment-by-visit interaction, as well as the continuous, fixed covariate of baseline value. For analyses of variables other than AHI, the AHI stratum will also be included in the model. Significance tests will be based on least squares means and Type III tests.

For continuous measures, summary statistics may include sample size, mean, standard deviation, median, minimum, and maximum for both the actual and the change from baseline measurements. Least-square means and standard errors derived from the analysis models will also be displayed for the change from baseline measurements. Treatment comparisons will be displayed showing the treatment difference least-square means and the 95% confidence intervals for the treatment differences, along with the p-values for the treatment comparisons.

For categorical measures, summary statistics may include sample size, frequency, and percentages. Fisher’s exact test or Pearson’s chi-square test will be used for treatment comparisons unless otherwise specified.

Not all analyses described in this SAP will necessarily be included in the CSRs. Any analysis described in this SAP and not provided in the CSR would be available upon request.

4.1.1. Adjustment for Covariates

The study is stratified by country/geographic region, OSA severity (not severe [AHI <30], severe [AHI ≥30]), and gender. Unless otherwise specified, the following factors will be adjusted for: geographic region (US/OUS), OSA severity (not severe [AHI <30], severe [AHI ≥30]), and gender. The value for stratification factors will be obtained from the data collected or derived from the eCRF or PSG results. In addition, the baseline value of the endpoint will be used as a covariate when appropriate.

4.1.2. Handling of Dropouts or Missing Data

For the primary and key secondary efficacy endpoint analyses aligned to the treatment regimen estimand and subject to Type 1 error rate control, missing data will be imputed based on the reason for the missing values, as described in Table GPIF.4.2. For analyses aligned to the "efficacy" estimand, missing data will be considered missing at random and hence no explicit imputation will be performed.

For exploratory endpoints and safety analyses, missing values will not be explicitly imputed unless specified otherwise.

For analyses aligned to the treatment regimen estimand, the statistical inference over multiple imputations will be guided by the method proposed by Rubin (1987). The missing values will be handled as follows:

Table GPIF.4.2. Imputation Approaches to Handle Missing/Invalid Data for Treatment Regimen Estimand

Missing/Invalid Data	Strategy to Handle Missing/Invalid Data	Assumptions for Missing Values	Methods to Handle Missing Values
Data missing at baseline, invalid data collected or missing data after treatment DC due to the COVID-19 pandemic (after other reasons for missing data are ruled out), technical issues (that is, sensor error on PSG) leading to invalid measurements ascertained while on treatment, or missing data after study DC due to inadvertent enrollment.	Hypothetical	MAR	Multiple imputation assuming MAR
Missing data due to any other reason (for example, study DC due to any reason other than COVID-19 or inadvertent enrollment).	Treatment policy	MNAR	Retrieved dropout imputation[a]. If there are not enough retrieved dropouts to provide a reliable imputation model, placebo-based multiple imputation will be used.

Abbreviations: COVID-19 = coronavirus disease-2019; DC = discontinuation; MAR = missing at random; MNAR = missing not at random; PSG = polysomnography; SBP = systolic blood pressure.

[a] Retrieved dropout imputation utilizes observed data from participants in the same treatment group who had outcome measures at Week 52 (or Week 48 for SBP) after early DC of study drug to impute the missing value.

4.1.3. Multicenter Studies

Randomization will be stratified by country, and geographic region (US/OUS) will be used as a covariate.

4.1.4. Historical Illnesses and Preexisting Conditions

The count and percentages of participants with historical illnesses and preexisting conditions will be summarized by treatment group using the MedDRA PTs nested within SOC. The SOC will be in alphabetical order. Conditions (that is, PTs) will be ordered by decreasing frequency in the tirzepatide MTD arm within the SOC. This will be summarized for all randomized participants.

4.1.5. Participant Characteristics

A listing of participant demographics for all randomized participants will be provided. The demographic and baseline clinical characteristics will also be summarized by study treatment for all randomized participants. Baseline demographic and clinical characteristics of special interest include but are not limited to:

- age (years)
- sex (female, male)
- race
- ethnicity
- height (cm)
- weight (kg)
- BMI (kg/m^2)
- waist circumference (cm)
- age group (<50, ≥50)
- BMI group (<35, ≥35 and <40, ≥40 kg/m^2)
- OSA severity (none [AHI <5], mild [AHI ≥5 and AHI <15], moderate [AHI ≥15 and AHI <30], or severe [AHI ≥30])
- geographic region (US/OUS), and
- country.

4.1.6. Concomitant Therapy

Concomitant medication will be summarized by treatment groups and displayed by decreasing frequency of WHODrug PTs in tirzepatide MTD arm. Baseline use of concomitant medication is defined as any medication started prior to the treatment start date and continuing on or after the treatment start date. Postbaseline concomitant medications are defined as those that are being taken any time during the postbaseline period.

In addition, medications of interest (as defined below) will be summarized by treatment groups:

- baseline use of:
    - lipid lowering therapy, by type/class and
    - antihypertensive therapy, by type/class

- Utilization after randomization of
  - antihyperglycemic medication for the treatment of diabetes for participants who develop type 2 diabetes mellitus during the study
  - antidiarrheal medication, and
  - antiemetic medication.

In addition, for ISA2 participants only, a summary of PAP machine use at baseline and postbaseline PAP machine adherence will be provided. Further details are provided in Section 4.7.1. For ISA1 participants, a listing will be provided summarizing any participants who use a PAP machine during the course of the trial.

4.1.7. Treatment Exposure and Compliance

4.1.7.1. Study and Study Treatment Exposure

Summary of duration of follow-up (defined as time in days from date of randomization to the date of the last study visit) will be provided by treatment group in the mITT population. Summary of duration on study treatment (defined as time in days from date of first dose of study treatment to date of last dose of study treatment plus 7 days) will be provided by treatment group in the safety analysis set.

For the summary of duration on study treatment, the frequency and percentage of participants falling into the following categorical ranges will also be summarized by planned treatment group as well: >0 week, ≥4 weeks, ≥8 weeks, ≥12 weeks, ≥16 weeks, ≥20 weeks, ≥24 weeks, ≥36 weeks, ≥48 weeks, and ≥52 weeks.

No p-values will be reported in these summaries as they are intended to describe the study populations rather than test hypotheses.

4.1.7.2. Adherence to Study Treatment

Summary of prematurely discontinuing study treatment (including reason for discontinuation) will be provided by study treatment. A time-to-event analysis of premature study treatment discontinuation will also be conducted.

If data warrants, the counts and percentages of participants who follow the planned escalation scheme, have dose interruption, or have dose de-escalation will be summarized for the tirzepatide treatment group. This will include the percentage of participants who have 10 mg or 15 mg tirzepatide as their MTD. In addition, the proportion of participants receiving 2.5, 5, 7.5, 10, 12.5, or 15 mg may be presented by randomized tirzepatide treatment and visit during the dose escalation period.

Treatment adherence will be defined as taking at least 75% of the scheduled tirzepatide doses. Treatment adherence will be summarized descriptively over the treatment period by treatment using the mITT population.

4.1.8. Important Protocol Deviations

Important protocol deviations are identified in the Trial Issues Management Plan. A listing and a summary of important protocol deviations by treatment will be provided.

4.2. Participant Dispositions

The participant dispositions for the screening period, the study intervention/treatment period, and/or the follow-up period will be collected in CRFs with the corresponding primary reason. The study completion for a participant is defined as the participant completing both the treatment period and the follow-up period, regardless of completion of study treatment.

Summaries and a listing of study disposition and study drug disposition will be provided for all randomized participants, separately for each ISA. Comparison between treatment arms will be performed using Fisher's exact test.

4.3. Primary Endpoint Analysis

The primary objective of this study is to test the hypothesis that tirzepatide at the MTD (10 mg or 15 mg) is superior to placebo for participants with moderate to severe OSA and obesity on the mean AHI reduction from baseline to Week 52. The primary and key secondary efficacy analyses will be guided by 2 estimands, the "treatment regimen" estimand and the "efficacy" estimand to support global regulatory submissions and publications.

4.3.1. Analysis Related to the Efficacy Estimand

The primary analysis guided by the "efficacy" estimand will be conducted using the EAS. This analysis will be based on the contrast between tirzepatide at the MTD (10 mg or 15 mg) and placebo at Week 52 (Visit 11) from the MMRM analysis of mean change from baseline in AHI. All the longitudinal observations at each scheduled postbaseline visit will be included in the analysis. REML will be used to obtain model parameter estimates and the Kenward-Roger option will be used to estimate the denominator degrees of freedom. The response variable of the MMRM will be the change in AHI from baseline values obtained at each scheduled postbaseline AHI measurement.

The model will include the fixed class effects of treatment, strata (geographic region [US/OUS] and gender), visit, and treatment-by-visit interaction, as well as the continuous, fixed covariate of baseline AHI. An unstructured covariance structure will be used to model the within-participant errors. Significance tests will be based on least squares means and Type III tests. If this analysis fails to converge, the following covariance structures will be tested in order until convergence is achieved:

- Toeplitz with heterogeneity
- autoregressive with heterogeneity
- compound symmetry with heterogeneous variances
- Toeplitz
- autoregressive, and
- compound symmetry without heterogeneous variances.

Analysis aligned to each estimand will be evaluated at the full significance level of 0.05.

4.3.2. Analysis Related to the Treatment Regimen Estimand

For the primary analysis guided by the "treatment regimen" estimand, the analysis will be conducted using the FAS. Missing values will be imputed based on the strategy to handle intercurrent events described in Section 4.1.2. After imputation, the primary efficacy comparison will be based on the contrast between tirzepatide at the MTD (10 mg or 15 mg) and placebo from the ANCOVA analysis of mean change from baseline to Week 52 in AHI using FAS. The ANCOVA model will include treatment and strata (geographic region [US/OUS] and gender) as fixed effects and baseline AHI as a fixed covariate. Statistical inference over multiple imputed data sets will be guided by Rubin (1987).

4.3.3. Sensitivity Analyses

For participants in ISA2, a sensitivity analysis will be carried out for the primary endpoint. When carrying out this sensitivity analysis using a treatment regimen estimand, participants with PAP withdrawal less than 5 days before the PSG at baseline or at Week 52 will have their data censored. Censored postbaseline data will be imputed using the approach outlined in Section 4.1.2. An ANCOVA model will be fit using the approaches outlined in Section 4.3.2. Carrying out this sensitivity analysis using the efficacy estimand, observations made with PAP withdrawal less than 5 days prior to PSG from the MMRM will be censored. Additional sensitivity analyses for ISA2 participants to accommodate participants with PAP withdrawal <5 days prior to the PSG may be considered.

Additional sensitivity analyses for both ISAs may be included as needed.

4.4. Secondary Endpoints Analysis

4.4.1. Key Secondary Endpoints

A graphical approach for multiple comparisons will be used to strongly control the overall Type 1 error (2-sided alpha level of 0.05) for testing the superior treatment effect of tirzepatide MTD over placebo including the key secondary endpoints as listed below.

- percent change in AHI at Week 52
- percent of participants with ≥50% AHI reduction at Week 52
- percent of participants at Week 52 with
  - AHI <5 or
  - (AHI 5 through 14 and ESS ≤10)
- percent change from baseline to Week 52 in body weight
- change from baseline to Week 48 in SBP
- change from baseline to Week 52 in C-reactive protein (high-sensitivity C-reactive protein)
- change in sleep apnea-specific hypoxic burden (% minutes/hour)
- change in PROMIS Sleep-related impairment short form 8a, and
- change in PROMIS Sleep disturbance short form 8b.

Analytical approaches for the hierarchical assessment of PROs are described in Section 4.4.3.1 and a summary of the analysis approach for all other key secondary endpoints is provided in Section 4.4.1.1.

4.4.1.1. Main Analytical Approaches

Analysis of percent change in AHI, percent change from baseline to Week 52 in body weight, change from baseline to Week 52 in log of high-sensitivity C-reactive protein, change from baseline to Week 48 in SBP, change from baseline to Week 52 in PROMIS Sleep-related impairment, change from baseline to Week 52 in PROMIS Sleep disturbance, and change from baseline to Week 52 in log of hypoxic burden will be conducted in a manner similar to the primary efficacy analyses using an ANCOVA model with treatment, strata (geographic region [US/OUS], AHI stratum [not severe (AHI <30), severe (AHI ≥30)], and gender), and baseline of the corresponding variable as a covariate for the treatment regimen estimand. If the hypoxic burden is reported to be 0, log (1) will be used in place of the log of hypoxic burden. The analysis method utilizing data from both ISAs for change from baseline to Week 52 in PROMIS sleep Impairment and PROMIS Sleep disturbance is described in the integrated efficacy analysis plan.

For the efficacy estimand, the MMRM analyses will be conducted as described in Section 4.1. For both estimands, analysis of percent change in AHI will adjust for the continuous, fixed baseline value of AHI instead of the baseline AHI stratum (not severe, severe).

Comparisons at the 52-week visit between the treatments relative to the proportion of participants achieving ≥50% AHI reduction and AHI<5 or (AHI 5 through 14 and ESS ≤10) will be conducted using logistic regression analysis including the following terms as a covariate:

- treatment
- geographic region (US/OUS)
- baseline AHI, and
- gender.

Unconditional risk differences will also be provided for these endpoints using logistic regression (Ye et al. 2023).

Analysis aligned to each estimand will be evaluated at the full significance level of 0.05 contingent on reaching statistical significance of the primary objective.

4.4.1.2. Sensitivity Analyses

For participants in ISA2, sensitivity analyses for the key secondary endpoints will be carried out: percent change in AHI, clinically meaningful change in AHI, and achieving OSA remission/mild nonsymptomatic OSA.

For percent change in AHI, the sensitivity analysis using both the treatment regimen and efficacy estimand will be carried out. When using a treatment regimen estimand approach, participants with PAP withdrawal less than 5 days before the PSG at baseline or at Week 52 will have their data censored. Censored postbaseline data will be imputed using the approach outlined in

Section 4.1.2. An ANCOVA model will be fit using the approaches outlined in Section 4.3.2. When using the efficacy estimand, observations made with a PAP withdrawal less than 5 days prior to the PSG or PROMIS from the MMRM will be censored.

For the binary endpoints of clinically meaningful change in AHI and achievement of OSA remission/mild nonsymptomatic OSA, the sensitivity analysis using both estimands will be carried out. PSG measurements taken after <5 days of PAP withdrawal will be censored. After censoring, analysis will be carried out as described in Section 4.4.1.1.

Additional sensitivity analyses for ISA2 participants to accommodate participants with PAP withdrawal <5 days prior to the PSG or PROMIS or ESS may be considered.

A 2-way tipping point analysis may also be utilized for the primary endpoint. This analysis will begin with the primary analysis aligned to the treatment regimen estimand and then adding positive and negative penalties simultaneously to both the tirzepatide MTD arm and the placebo arm, considering when results tip from superiority to inconclusive, and then considering the clinical plausibility of such scenarios.

Additional sensitivity analyses for both ISAs may be included as needed.

4.4.2. Type 1 Error Rate Control Strategy for Primary and Key Secondary Efficacy Analyses

All primary and key secondary hypotheses will be tested with the overall family-wise Type 1 error rate at a 2-sided alpha level of 0.05 through the multiplicity control approach based on the graphical multiple testing procedure. The primary endpoint hypothesis will be tested at a 2-sided alpha level of 0.05 for statistical significance. If the primary efficacy endpoint is significant, the alpha of 0.05 will be propagated to the key secondary efficacy endpoints. The detailed graphical testing scheme is outlined in Figure GPIF.2.1.

The analyses will be performed for both the treatment regimen and efficacy estimands described in Section 4.3 using the same graphical testing scheme. An overall 2-sided alpha of 0.05 to control Type 1 error rate separately for the treatment regimen estimand and the efficacy estimand will be used.

4.4.3. Supportive Secondary Endpoints

Unless otherwise specified, all supportive/other secondary efficacy analyses will be guided by the "efficacy" estimand and will be conducted using the EAS. Missing data will be handled using an MMRM.

Table GPIF.4.3. Secondary Measures Not Controlled for Type 1 Error

Objective – Demonstrate Superiority of Tirzepatide MTD to Placebo for:	Endpoint	Analytical Approaches
Change in excessive daytime sleepiness	Change in ESS score from baseline to Week 52	MMRM analysis described in Section 4.4.1.1 will be conducted.

Objective – Demonstrate Superiority of Tirzepatide MTD to Placebo for:	**Endpoint**	**Analytical Approaches**
Change in patient-reported functional status as assessed by FOSQ (10 items)	Change in FOSQ-10 total score from baseline to Week 52	MMRM analysis described in Section 4.4.1.1 will be conducted.
Change in patient-reported functional status as assessed by FOSQ (30 items)	Change in FOSQ (30 item) total score and all functional domain scores from baseline to Week 52	MMRM analysis described in Section 4.4.1.1 will be conducted.
Change in Body Weight	Percent of participants who achieve ≥10%, ≥15%, and ≥20% body weight reduction.	Logistic models described in Section 4.4.1.1 with the following covariates: treatment, geographic region (US/OUS), baseline AHI (not severe/severe), gender, and baseline bodyweight as a covariate.
Change in Lipid Parameters	Change in: HDL-cholesterol non-HDL-cholesterol triglycerides	MMRM analysis described in Section 4.4.1.1 will be conducted.
Hierarchical assessment of PRO change	A hierarchical combination of the following: ○ Change in PROMIS Sleep-related impairment short form 8a ○ Change in PROMIS Sleep disturbance short form 8b	Win ratio analysis described in Section 4.4.3.1 will be conducted.
Change in supportive secondary PROs	Summary of item 8 of PROMIS Sleep Disturbance short form 8b	Counts and percentages of participants at each time point will be summarized by nominal visit and by treatment. A shift table from baseline to postbaseline will be created at each postbaseline visit.
	Change in: SF-36v2 acute form domain scores From baseline to Week 52	MMRM analysis of T-score described in Section 4.4.1.1 will be conducted. Description of T-score calculation provided in Sections 4.7.2.4 and 4.7.2.5.
	Percent of participants with improved categorical shift in: PGIS-OSA Sleepiness PGIS-OSA Fatigue PGIS-OSA Snoring From baseline to Week 52	For each question, the proportion of participants with improvements from baseline will be summarized. Shift analysis from baseline to Week 52 will also be performed.
	Proportion of participants who achieve: $\leq -x$ change in PROMIS Sleep-related impairment $\leq -y$ change in PROMIS Sleep disturbance	Logistic models described in Section 4.4.1.1 with the following covariates: treatment, geographic region (US/OUS) baseline AHI (not severe/severe), gender, and baseline score as a covariate.

Objective – Demonstrate Superiority of Tirzepatide MTD to Placebo for:	Endpoint	Analytical Approaches
	From baseline to Week 52 (x and y will be derived from blinded interim analysis)	
Change in Insulin	Change in fasting insulin from baseline to Week 52	MMRM analysis described in Section 4.4.1.1 will be conducted.
Change in DBP	Change in DBP from baseline to Week 48	MMRM analysis described in Section 4.4.1.1 will be conducted.

Abbreviations: AHI = Apnea-Hypopnea Index; DBP = diastolic blood pressure; ESS = Epworth Sleepiness Scale; FOSQ = Functional Outcomes of Sleep Questionnaire; HDL = high-density lipoprotein; MMRM = mixed model repeated measures; MTD = maximum tolerated dose; PGIS-OSA = Patient Global Impression of Status Obstructive Sleep Apnea; PRO = patient-reported outcome; PROMIS = Patient-Reported Outcomes Measurement Information System; SF-36v2 = Short-Form 36 version 2.

4.4.3.1. Hierarchical Assessment of PRO Change

The analysis of the hierarchical composite endpoint will be performed with the Finkelstein-Schoenfeld method, and the win ratio (Pocock et al. 2012) will be reported as the measure of treatment effect. The population level summary of win ratio will be calculated as number of pairs of tirzepatide-treated participant "wins" divided by number of pairs of placebo-treated participant "wins."

The Finkelstein-Schoenfeld method is based on the principle that each tirzepatide-treated participant is compared with every other placebo-treated participant in a pairwise manner that proceeds in a hierarchical fashion. Differences will be calculated as tirzepatide participant value minus placebo participant value. Each pairwise comparison will proceed in the following order and a winner has:

- Stage 1: For the change from baseline at Week 52 PROMIS Sleep-related impairment score:
  - A comparison is a win when the treatment difference is ≤-4.9
  - A comparison is a loss when the treatment difference is ≥4.9
  - All other cases are a tie and the comparison of PROMIS Sleep disturbance score will be conducted (that is, proceed to Stage 2).
- Stage 2: For the change from baseline at Week 52 PROMIS Sleep disturbance score:
  - A comparison is a win when the treatment difference is ≤-3.1.
  - A comparison is a loss when the treatment difference is ≥3.1.
  - In all other cases, the pair will be recorded as a tie.

Based on Donovan et al. (2020), a clinically important response of PROMIS sleep-related impairment for OSA participants is 4.9, and a clinically important response of PROMIS sleep disturbance for OSA participants is 3.1. Participants in this study had a mean BMI of 33.7 kg/m^2, and one-half of them had moderate to severe OSA. Thus, the meaningful change threshold can be generalized to participants in the tirzepatide OSA trial.

For treatment policy estimand, missing values at Week 52 will be imputed through multiple imputations based on the reason of missingness with details described in Section 4.1.2.

For PRO measures, missing baseline values will be assumed to be missing at random and may be imputed through multiple imputation methods.

4.5. Exploratory Endpoint Analyses

Unless otherwise specified, all exploratory efficacy analyses will be guided by the "efficacy" estimand and will be conducted using the EAS. Missing data will be handled using an MMRM.

The following efficacy analyses apply to both ISAs and will be carried out separately for each ISA.

Table GPIF.4.4. Exploratory Efficacy Analysis for Both ISAs

Objective – Demonstrate Superiority of Tirzepatide MTD to Placebo for:	Endpoint	Analytical Approaches
Change in exploratory PROs	Change from baseline to Week 52 in: EQ-5D-5L utility index EQ-VAS scores	MMRM analysis described in Section 4.4.1.1 will be conducted.
	Percent of participants with improved categorical shift from baseline to Week 52 in: PGIC-OSA Sleepiness PGIC-OSA Fatigue PGIC-OSA Sleep quality PGIC-OSA Snoring	For each question, the proportion of participants with improvements from baseline will be summarized. Shift analysis from baseline to Week 52 will also be performed.
Change in parameters measured by Actigraphy (AX6)	Mean change from baseline to Week 52 in: Daytime sleep duration Daily step counts Average acceleration	MMRM analysis described in Section 4.4.1.1 will be conducted.

Abbreviations: EQ-5D-5L = EuroQoL-5 Dimension-5 Level; EQ-VAS = EuroQol Visual Analogue Scale; ISA = intervention-specific appendix; MMRM = Mixed model repeated measures; MTD = maximum tolerated dose; PRO = patient-reported outcome; PGIC-OSA = Patient Global Impression of Change Obstructive Sleep Apnea.

The efficacy analyses summarized in Table GPIF.4.5 only apply to participants in ISA1.

Table GPIF.4.5. Exploratory Efficacy Analysis Conducted only for ISA1 Participants

Objective – Demonstrate Superiority of Tirzepatide MTD to placebo for:	Endpoint	Analytical Approaches
Change in parameters measured by WatchPAT300	Change from baseline to Week 52 in PAT-based device determinations of: • pAHI • SASHB	MMRM analysis described in Section 4.4.1.1 will be conducted.

Abbreviations: MMRM = mixed model repeated measures; MTD = maximum tolerated dose; pAHI = peripheral tone apnea-hypopnea index; PAT = peripheral arterial tonometry; SASHB = sleep apnea specific hypoxic burden.

4.6. Safety Analyses

Unless specified otherwise, safety assessments will be guided by the safety estimand. Thus, unless specified otherwise, safety analyses will be conducted in the safety analysis set (Table GPIF.3.1); all events that occur between the first dose date of study drug and the end date of study participation will be included, regardless of the adherence to study drug.

The statistical assessment of homogeneity of the distribution of categorical safety responses between tirzepatide MTD and placebo will be conducted using Fisher's exact test, unless specified otherwise.

The mean change from baseline differences among treatments at all scheduled visits will be assessed via an MMRM using REML. The model will include treatment group, stratification factors, visit and treatment-by-visit interaction as fixed effects, and baseline value of the safety parameter as a covariate. To model the covariance structure within participants, the unstructured covariance matrix will be used. If this model fails to converge, the covariance structures specified in Section 4.3.1 will be tested in order until convergence is met. If the data does not warrant the MMRM model, then an ANCOVA model will be used.

For selected safety parameters, time-to-first-event analysis via the Cox-proportional hazards model may be conducted. Participants without the event will be censored at the end of study participation. For participants experiencing the event, the "time-to-first-event" will be the time (in days) from first dose to first occurrence of the event.

4.6.1. Analysis of Adverse Events

4.6.1.1. Treatment-Emergent Adverse Events

A TEAE is defined as an event that first occurred or worsened in severity after the first dose of study treatment. The MedDRA LLT will be used in the treatment-emergent derivation. The maximum severity for each LLT during the baseline period including ongoing medical history will be used as baseline severity. For events with a missing severity during the baseline period, it will be treated as "mild" in severity for determining treatment-emergence. Events with a missing severity during the postbaseline period will be treated as "severe" and treatment-emergence will be determined by comparing to baseline severity.

For events occurring on the day of taking study medication for the first time, the CRF-collected information (for example, if the event starts or worsens after the first dose) will be used to determine whether the event was pre- versus posttreatment if available. If the relevant information is not available, then the events will be counted as posttreatment.

Unless otherwise specified, the counts and percentages of participants with TEAEs will be summarized by treatment using MedDRA PT nested within SOC. Statistical comparisons will be applied at both the SOC and PT levels. Events will be ordered by decreasing frequency in the tirzepatide arm within the SOC. The SOC will be in alphabetical order.

An overview of the number and percentage of participants who experienced a TEAE, SAE, death, discontinued from study treatment or study due to an AE, or with a TEAE related to study treatment will be summarized by treatment.

The counts and percentages of participants with TEAEs by maximum severity will be summarized by treatment using the MedDRA PT within the SOC. For each participant and TEAE, the maximum severity for the MedDRA PT is the maximum postbaseline severity observed from all associated LLTs mapping to the MedDRA PT. The maximum severity will be determined based on the nonmissing severities. If all severities are missing for the defined postbaseline period of interest, it will show as missing in the table.

For events that are gender specific, the denominator and computation of the percentage will only include participants of the given gender.

4.6.1.2. Common Adverse Events

The counts and percentages of participants with TEAEs, overall and common (common TEAEs occurred in ≥5% of participants before rounding), will be summarized by treatment using MedDRA PT. Events will be ordered by decreasing frequency in the tirzepatide arm.

4.6.1.3. Deaths

A listing of all deaths during the study will be provided. The listing will include participant identification including:

- treatment
- site number
- date of death
- age at the time of enrollment
- sex
- associated AE group identification
- time from last dose of study drug to death (if participant had discontinued study drug), and
- primary cause of death.

4.6.1.4. Other Serious Adverse Events

The counts and percentages of participants who experienced an SAE (including deaths and SAEs temporally associated or preceding deaths) during the postbaseline period will be summarized by treatment using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the tirzepatide arm within the SOC. The SOC will be in alphabetical order.

A listing of all SAEs will be provided. The listing will include:

- treatment
- participant identification including the site number
- date of event
- age at the time of enrollment
- sex

- AE group identification
- MedDRA SOC and PT
- severity
- outcome
- relationship to study drug
- time from first dose of study drug to the event, and
- time from most recent dose to event (if participant discontinued study drug prior to the event).

4.6.1.5. Discontinuation Due to Adverse Events

The counts and percentages of participants who discontinued from study treatment or study due to an AE during the postbaseline period may be summarized by treatment group using MedDRA PT nested within SOC. Events will be ordered by decreasing frequency in the tirzepatide arm within the SOC.

4.6.2. Patient Narratives

Patient narratives will be provided for all participants who experience any of the following "notable" events:

- death
- SAE
- pregnancy, or
- permanent discontinuation of study treatment due to AEs.

Patient narratives (patient level data and summary paragraph) will be provided for participants in the randomized population with at least 1 notable event.

4.6.3. Special Safety Topics

For AESI or special safety topics, the counts and percentages of participants will be summarized by treatment and PT with decreasing frequency in the tirzepatide arm if the overall count is 10 or more. Individual participant level data may be presented. Displays with individual participant level data may be created using various formats, such as a customized listing and/or a customized graphical participant profile. AESI are defined in each section of special safety topics, where applicable.

4.6.3.1. Exocrine Pancreas Safety

4.6.3.1.1. Pancreatic Enzyme

Observed pancreatic enzyme data (p-amylase and lipase) will be summarized by treatment and nominal visit.

The counts and percentages of participants with maximum postbaseline pancreatic enzyme value exceeding the following thresholds will be provided by maximum baseline pancreatic enzyme value (≤ULN, >ULN), and postbaseline:

- ≤1 × ULN
- (>1 to ≤3) × ULN

- (>3 to ≤5) × ULN
- (>5 to ≤10) × ULN, and
- >10 × ULN.

An MMRM analysis will be used to analyze each pancreatic enzyme with a log-transformed (postbaseline measure/baseline measure) response variable and treatment, nominal visit, treatment-by-nominal visit interaction as fixed effects.

4.6.3.1.2. Pancreatitis Events

Summaries of adjudicated and investigator-reported pancreatic events will be provided by treatment. Detailed searching criteria can be found in Appendix 2 (Section 6.2).

Treatment-emergent adjudication-confirmed pancreatitis will be considered as an AESI. Listing of participants with adjudicated pancreatitis may be provided if deemed necessary.

4.6.3.2. Gastrointestinal Adverse Events

4.6.3.2.1. Nausea, Vomiting, and Diarrhea

Summaries and analyses for incidence and severity of nausea, vomiting (including "vomiting" and "vomiting projectile"), diarrhea (including "diarrhea" and "diarrhoea"), and 3 events combined, will be provided by each treatment group.

Summary of the prevalence over time for nausea, vomiting, and diarrhea will also be presented. Time to the onset of nausea, vomiting, and diarrhea will be plotted.

4.6.3.2.2. Severe Gastrointestinal Events

The PTs under the *Gastrointestinal disorders* SOC in MedDRA will be used to identify GI AEs, and only the PTs with serious/severe treatment-emergent cases will be considered as AESIs.

The counts and percentages of participants with severe/serious treatment-emergent GI events may be summarized by treatment, or a listing may be provided.

4.6.3.3. Hepatobiliary Disorders

4.6.3.3.1. Hepatobiliary Events

The counts and percentages of participants with treatment-emergent hepatic events may be summarized by treatment using the MedDRA PTs. The detailed search criteria can be found in Appendix 2 (Section 6.2).

Events related to acute gallbladder disease may also be summarized or a listing may be provided. The search criteria can be found in Appendix 2 (Section 6.2).

Severe/serious treatment-emergent hepatic events and acute gallbladder disease will be considered as AESIs.

4.6.3.3.2. Liver Enzymes

Common analyses for laboratory analyte measurements described in Section 4.6.5 are applicable for the liver enzyme related measurements. This section describes additional analyses for liver enzymes.

For the postbaseline maximum value, all planned and unplanned measurements will be included. When or if multiple values are available (that is, unplanned measurement) prior to randomization, the maximum value will be used as baseline. Table GPIF.4.6 describes the planned analyses related to hepatic safety.

Table GPIF.4.6. Summary Tables and Figures Related to Hepatic Safety

Analysis	Population or Analysis Set

Abnormal Postbaseline Categories – Hepatic Safety Parameters
• ALT: The number and percentage of participants with a measurement greater than or equal to 1 time (1×), 3 times (3×), 5 times (5×), 10 times (10×), and 20 times (20×) the performing laboratory ULN during the treatment period for all participants with a postbaseline value and for subsets based on the following levels of baseline value:
  ○ participants whose nonmissing maximum baseline value is ≤1 × ULN,
  ○ participants whose maximum baseline is >1 × ULN,
  ○ participants whose baseline values are missing.
• AST: The number and percentage of participants with a measurement greater than or equal to 1 time (1×), 3 times (3×), 5 times (5×), 10 times (10×), and 20 times (20×) the performing laboratory ULN during the treatment period for all participants with a postbaseline value and for subsets based on various baseline levels, as described above for ALT.
• ALP: The number and percentage of participants with a measurement greater than or equal to 2 times (2×), and 3 times (3×) the performing laboratory ULN during the treatment period will be summarized for all participants with a postbaseline and for the following subsets based on the baseline values:
  ○ participants whose nonmissing maximum baseline value is ≤1 × ULN,
  ○ participants whose maximum baseline is >1 × ULN, but <2 × ULN,
  ○ participants whose maximum baseline value is ≥2 × ULN, and
  ○ participants whose baseline values are missing.
• TBL: The number and percentage of participants with a measurement greater than or equal to 2 times (2×), 5 times (5×), and 8 times (8×) the performing laboratory ULN during the treatment period will be summarized for all participants with a postbaseline value and the same subsets as described for ALP.
• DBL: The number and percentage of participants with a measurement greater than or equal to 2 times (2×) and 5 times (5×) the performing laboratory ULN during the treatment period will be summarized for all participants with a postbaseline value and the same subsets as described for ALP.
• GGT: The number and percentage of participants with a measurement greater than or equal to 2 times (2×) the performing laboratory ULN during the treatment period will be summarized for all participants with a postbaseline value.
	Safety Participants

Hepatocellular Drug-Induced Liver Injury Screening Plot (TBL vs. ALT or AST).	Safety Participants
Hepatocellular Drug-Induced Liver Injury Screening Table.	Safety Participants

Analysis	Population or Analysis Set
Cholestatic Drug-Induced Liver Injury Screening Plot (TBL vs. ALP).	Safety Participants
Cholestatic Drug-Induced Liver Injury Screening Table.	Safety Participants
Participant profiles will be created for participants meeting criteria for a comprehensive hepatic evaluation (as defined in the protocol). Participant profiles will include demographics, disposition, information collected on the hepatic safety CRFs (where applicable) and a display of study drug exposure, adverse events, medications, blood pressure, heart rate, and the liver -related measurements over time.	Safety Participants

Abbreviations: ALP = alkaline phosphatase; ALT = alanine aminotransferase; AST = aspartate aminotransferase; CRF = case report form; DBL = direct bilirubin; GGT = gamma-glutamyl transferase; TBL = total bilirubin; ULN = upper limit of normal.

4.6.3.4. Hypoglycemia

The following categories in accordance with the 2020 American Diabetes Association position statement on glycemic targets (ADA 2020) will be defined in the database.

Level 1 hypoglycemia

Glucose <70 mg/dL (3.9 mmol/L) and ≥54 mg/dL (3.0 mmol/L): Level 1 hypoglycemia can alert a person to take action such as treatment with fast-acting carbohydrates. Providers should continue to counsel participants to treat hypoglycemia at this glucose alert value.

Level 2 hypoglycemia

Glucose <54 mg/dL (3.0 mmol/L): Level 2 hypoglycemia is also referred to as documented or BG confirmed hypoglycemia with glucose <54 mg/dL (3.0 mmol/L). This glucose threshold is clinically relevant regardless of the presence or absence of symptoms of hypoglycemia.

Level 3 hypoglycemia

Severe hypoglycemia (in adults): A severe event characterized by altered mental and/or physical status requiring assistance for treatment of hypoglycemia. For example, participants had altered mental status, and could not assist in their own care, or were semiconscious or unconscious, or experienced coma with or without seizures, and the assistance of another person was needed to actively administer carbohydrate, glucagon, or other resuscitative actions. Glucose measurements may not be available during such an event, but neurological recovery attributable to the restoration of glucose concentration to normal is considered sufficient evidence that the event was induced by a low glucose concentration.

- The determination of a hypoglycemic event as an episode of severe hypoglycemia, as defined above, is made by the investigator based on the medical need of the participant to have required assistance and is not predicated on the report of a participant simply having received assistance.

- If a hypoglycemic event meets the criteria of severe hypoglycemia, the investigator must record the event as serious on the AE CRF and report it to Lilly as an SAE.

Nocturnal hypoglycemia

Nocturnal hypoglycemia is a hypoglycemia event (including severe hypoglycemia) that **occurs at night** and presumably during sleep.

To avoid duplicate reporting, all consecutive BG values <70 mg/dL (<3.9 mmol/L) occurring within a 1-hour period may be considered to be a single hypoglycemic event (Weinberg et al. 2010; Danne et al. 2013).

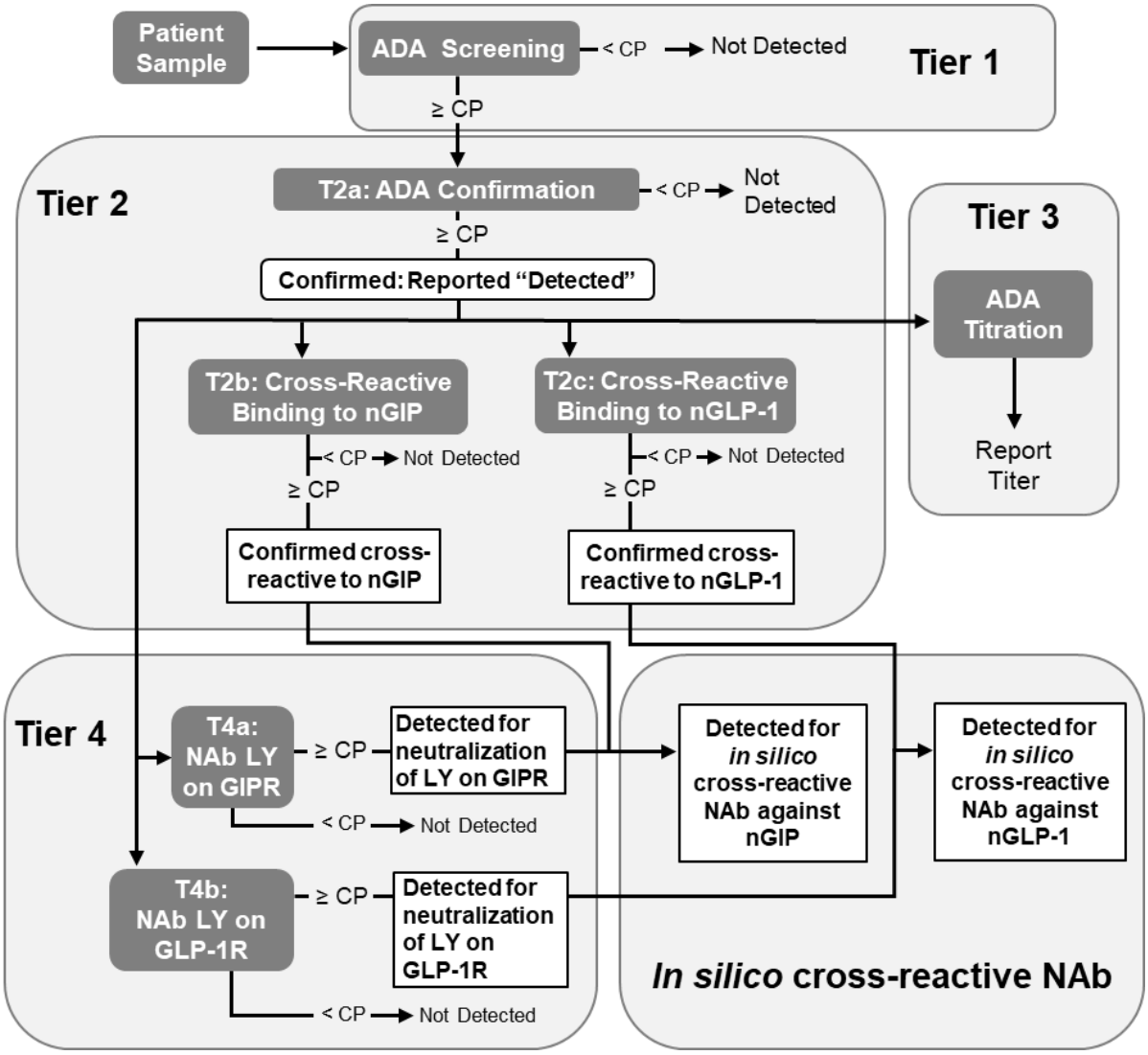
Summary and analyses of Level 2 or Level 3 hypoglycemic events will be performed.

4.6.3.5. Immunogenicity

4.6.3.5.1. Definitions of Sample ADA Status

At a high level, an individual sample is potentially examined multiple times, in a hierarchical procedure, to produce a sample ADA assay result and potentially multiple cross-reactive antibodies assay results and multiple Nab assay results.

The cut points used, the drug tolerance of each assay, and the possible values of titers are operating characteristics of the assay. Figure GPIF.4.1 details a flow chart that reflects the multitiered testing approach.



Abbreviations: ADA = anti-drug antibody; CP = cut point; GIP = glucose-dependent insulinotropic polypeptide; GIPR = glucose-dependent insulinotropic polypeptide receptor; GLP-1 = glucagon-like peptide-1; GLP-1R = glucagon-like peptide-1 receptor; nGIP = native GIP; nGLP-1 = native GLP-1; LY = LY3298176; Nab = neutralizing antibodies.

Figure GPIF.4.1. Flowchart of immunogenicity multitiered testing approach.

Table GPIF.4.7 outlines results as reported from Tier 2a of the multitiered testing approach. Tier 4 results are reported similarly.

Table GPIF.4.7. Sample ADA Assay Results

Sample Laboratory Result	Explanation
Detected	ADA are detected and confirmed.
Not Detected	The raw result as reported from the laboratory indicates not detected. The clinical interpretation of such results depends on other factors (see Table GPIF.4.8).
NO TEST, QNS, and so on.	Sample exists but was unevaluable by the assay.

Abbreviations: ADA = anti-drug antibody; QNS = quantity not sufficient.

It can be the case that the presence of high concentrations of tirzepatide will affect ADA immunoassays, and conversely high levels of ADA may affect the measurement of tirzepatide concentration. Thus, a tirzepatide drug concentration, assessed from a sample drawn at the same time as the ADA sample, plays a key role in clinical interpretation of a sample when the laboratory result is Not Detected (see Table GPIF.4.8).

Table GPIF.4.8. Sample Clinical ADA Interpretation Results

Sample Clinical Interpretation	Explanation
ADA Present	ADA assay result is Detected
ADA Not Present	ADA assay result is Not Detected and simultaneous drug concentration is at a level that has been demonstrated to not interfere in the ADA detection method (i.e., drug concentration is below the assay's drug tolerance level). For participants receiving placebo, drug concentration is not assessed and is assumed to be below the assay's drug tolerance level. If drug concentration was planned but is not available for a treatment period sample, a Not Detected sample will be declared ADA Not Present.
ADA Inconclusive	ADA assay result is Not Detected but drug concentration in the sample is at a level that can cause interference in the ADA detection method.
ADA Missing	ADA sample not drawn, QNS, not tested, and so on, causing there to be no laboratory result reported or the result is reported as "no test."

Abbreviations: ADA = antidrug antibody; QNS = quantity not sufficient.

All ADA Present samples will be evaluated for cross-reactive GIP (Tier 2b), cross-reactive GLP-1 (Tier 2c), Nab LY (tirzepatide) on GIPR (Tier 4a), and Nab LY (tirzepatide) on GLP-1R (Tier 4b).

Similar terminology to Table GPIF.4.8 applies for each type of cross-reactive and Nab assay. Importantly, each of these are distinct assays and, in general, have different assay operating characteristics. The following are considered inconclusive for the Nab result:

- Nab LY on GIPR: if Nab result is not detected, and PK concentration is greater than or equal to drug tolerance limit of the Nab LY on GIPR assay
- Nab LY on GLP-1R: if Nab result is not detected, and PK concentration is greater than or equal to drug tolerance limit of the Nab LY on GLP-1R assay

For cross-reactive Nab interpretations against native GIP and GLP-1, an *in silico* method utilizing results from Tiers 2b and 2c, Tiers 4a and 4b, and tirzepatide concentrations is introduced. The *in silico* method is outlined in Table GPIF.4.9.

Table GPIF.4.9. In Silico Classification for Cross-Reactive Nab

In Silico Classification	Cross-Reactive ADA Result	Nab Result	Circulating Tirzepatide Level (ng/mL)	*In Silico* Cross-Reactive Nab Interpretation
Cross-reactive Nab to nGIP	Tier 2b: "Not Detected"	Tier 4a: "Not Detected" Or Tier 4a: "Detected" or N/A or Missing	Any Value or Missing	Not Present
	Tier 2b: "Detected"	Tier 4a: "Not Detected"	< drug tolerance limit of Tier 4a assay	Not Present
	Tier 2b: "Detected"	Tier 4a: "Not Detected"	≥ drug tolerance limit of Tier 4a assay	Inconclusive
	Tier 2b: "Detected"	Tier 4a: "Detected"	< drug tolerance limit of Tier 4a assay	Present
	Tier 2b: "Detected"	Tier 4a: "Detected"	≥ drug tolerance limit of Tier 4a assay	Present
Cross-reactive Nab to nGLP-1	Tier 2c: "Not Detected"	Tier 4b: "Not Detected" Or Tier 4b: "Detected" or N/A or Missing	Any Value or Missing	Not Present
	Tier 2c: "Detected"	Tier 4b: "Not Detected"	< drug tolerance limit of Tier 4b assay	Not Present
	Tier 2c: "Detected"	Tier 4b: "Not Detected"	≥ drug tolerance limit of Tier 4b assay	Inconclusive
	Tier 2c: "Detected"	Tier 4b: "Detected"	< drug tolerance limit of Tier 4b assay	Present
	Tier 2c: "Detected"	Tier 4b: "Detected"	≥ drug tolerance limit of Tier 4b assay	Present

Abbreviations: ADA = antidrug antibody; Nab = neutralizing antibody; nGIP = native glucose-dependent insulinotropic polypeptide; nGLP-1 = native glucagon-like peptide-1; Tier 2b = cross-reactive ADA to nGIP; Tier 2c = cross-reactive ADA to nGLP-1; Tier 4a = Nab LY (tirzepatide) on GIPR; Tier 4b = Nab LY (tirzepatide) on GLP-1R.

Note: Only the drug tolerance limits of the Tier 4a and 4b assays are used for in silico classifications as they are lower than the drug tolerance limits of the Tier 2b and 2c assays, respectively.

4.6.3.5.2. Definitions of Immunogenicity Assessment Periods

Immunogenicity baseline observations: Baseline period for immunogenicity assessment for each participant includes all observations prior to first dose of study treatment. In instances where multiple baseline observations are collected, to determine participant ADA status the last nonmissing immunogenicity assessment prior to first administration of study drug is used to determine treatment-emergent status (see below).

Immunogenicity postbaseline period observations: Postbaseline period observations for each participant includes all observations after the first administration of study drug.

4.6.3.5.3. Definitions of Participant ADA Status

TE ADA-evaluable participants: A participant with a nonmissing baseline ADA result and at least 1 nonmissing postbaseline ADA result.

TE ADA-unevaluable participant: any participant who does not meet the evaluable criteria.

Baseline ADA Present (preexisting antibody): ADA detected in a sample collected up to the first dose date and time.

Baseline ADA Not Present: ADA is not detected, and the corresponding PK concentration is missing or below the drug tolerance limit in a sample collected up to the first dose date and time.

TE ADA+ participant: An evaluable participant who had a:

- baseline status of ADA Not Present and at least 1 postbaseline status of ADA Present with titer ≥2 × MRD, where the MRD is the minimum required dilution of the ADA assay or
- baseline and postbaseline status of ADA Present, with the postbaseline titer being 2 dilutions (4-fold) greater than the baseline titer. That is, the participant has baseline (B) status of ADA Present, with titer 1:B, and at least 1 postbaseline (P) status of ADA Present, with titer 1:P, with P/B ≥4.

As shown in Figure GPIF.4.1, a titer is expected when ADA assay result is Detected. On occasion, the corresponding assay cannot be performed, in which case a titer value will be imputed for the purpose of TE ADA determination. A baseline sample with detected ADA and no titer is imputed to be the MRD (1:10), and a postbaseline sample with ADA detected and no titer is imputed to be 1 dilution above the MRD (1:20).

TE ADA- Inconclusive participant: A TE ADA-evaluable participant is TE ADA Inconclusive if ≥20% of the participant's postbaseline samples, drawn predose, are ADA Inconclusive and all remaining postbaseline samples are ADA Not Present.

TE ADA- participant: A TE ADA-evaluable participant is TE ADA- when the participant is not TE ADA+ and not TE ADA Inconclusive.

For each Nab assay, the following are defined:

Nab+ participant: A participant who is TE ADA+ and has a Nab+ sample in the postbaseline period.

Nab Inconclusive participant: A participant who is TE ADA+, is not Nab+, and all samples that have TE ADA+ titer have a Nab Inconclusive sample result.

Nab- participant: A participant is neither Nab+ nor Nab Inconclusive.

Unless specified otherwise, the above-mentioned definitions of Nab are applicable to all Nab analyses, including cross-reactive Nab analyses, and cross-reactive antibodies.

4.6.3.5.4. Analyses to be Performed

The count and proportion of participants who are TE ADA+ will be tabulated by treatment group, where the proportions are relative to the number of TE ADA-evaluable participants, as defined above. The tabulation will include the count and proportion of participants with ADA Present at baseline, and the count and proportion of TE ADA+ participants exhibiting each type of cross-reactive antibodies and Nab. This analysis will be performed for the planned treatment

period. The cross-reactive Nab will include the *in silico* classification as cross-reactive Nab for summary.

Additional immunogenicity analyses as determined later may be presented. The relationship between the presence of antibodies and tirzepatide PK and pharmacodynamic response including safety and efficacy to tirzepatide may be assessed.

4.6.3.6. Hypersensitivity Reactions

Two main analyses are performed in support of assessment of potential immediate hypersensitivity, including anaphylaxis as well as potential nonimmediate hypersensitivity.

Time Period A, of potential immediate hypersensitivity includes all TEAEs occurring from start of study drug administration up to 24 hours after end of study drug administration. For events without the hypersensitivity eCRF, only date (no time) information is collected. Among these events without time information, the event occurred on the same date as the study drug injection date will be included in Time Period A.

Time Period B, of potential non-immediate hypersensitivity, includes all TEAEs occurring more than 24 hours after the end of study drug administration, but prior to subsequent drug administration.

Analyses for both time periods are based on the following:

- narrow and algorithm terms in *Anaphylactic reaction* SMQ (20000021) (analysis for algorithm term only applicable for Time Period A)
- narrow terms in *Angioedema* SMQ (20000024)
- narrow terms in *Severe cutaneous adverse reactions* SMQ (20000020), and
- narrow terms in *Hypersensitivity* SMQ (20000214)

For the *Anaphylactic reaction* SMQ, each term is classified by scope (narrow, broad) and by category (A, B, C, and D). All narrow terms are category A, and all broad terms are category B, C, or D. In addition to the usual narrow and broad searches, the SMQ defines an algorithm to further refine the cases of interest. For Time Period A analysis, the *Anaphylactic reaction* SMQ algorithm will be included. The algorithm is based upon events that occur within Time Period A. The counts and percentages of participants who experienced a TEAE for the following will be analyzed for each of the 2 time periods:

- any narrow term from any 1 of the 4 SMQs indicated above (that is, combined search across narrow of all 4 SMQs), and
- any narrow scope term within each SMQ, separately (that is, narrow SMQ search). For Time Period A analysis, any term from Anaphylactic reaction SMQ algorithm.

Within each query, individual PTs that satisfied the queries will be summarized. For Time Period A analysis, the Anaphylactic reaction SMQ algorithm will be summarized. Also, a single event may satisfy multiple SMQs, in which case the event contributes to every applicable SMQ.

4.6.3.6.1. Severe/Serious Hypersensitivity Reactions

The severe/serious cases of hypersensitivity will be considered as AESIs. Summary of severe/serious hypersensitivity reactions or listing may be provided.

4.6.3.7. Injection Site Reaction

Injection site reaction, incidence and rates, and related information reported via "Injection Site Reactions" eCRF will be summarized by treatment. Information to be summarized include location of the reaction, timing of reaction relative to study drug administration, and characteristics of the injection site reaction: erythema, induration, pain, pruritus, and edema.

Patient-based analysis and event-based analysis may be provided if necessary. The patient-based analysis summarizes all ISR questionnaire forms for an individual participant with a single statistic, typically an extreme value. This analysis allows each participant to contribute only once for each parameter, at the expense of a focus on the most extreme events. By contrast, the event-based analysis summarizes all ISR questionnaire forms received, without regard to individual participants. This provides characteristics of ISR events as a proportion of all events for which questionnaire responses were provided, at the expense of some potential bias due to differential contribution of individual participants to the analysis.

The counts and percentages of participants with treatment-emergent injection site reaction will be summarized by treatment using the MedDRA PTs. Detailed searching criteria can be found in Appendix 2 (Section 6.2).

The PTs will be summarized in decreasing order of incidence for tirzepatide-treated participants.

4.6.3.7.1. Severe/Serious Injection Site Reactions

Severe/serious injection site reactions (for example, abscess, cellulitis, erythema, hematomas/hemorrhage, exfoliation/necrosis, pain, subcutaneous nodules, swelling, induration, inflammation) will be considered as AESI.

The counts and percentage of participants with severe/serious ISRs may be summarized by treatment, or a listing of participants with treatment-emergent severe/serious ISRs may be provided.

4.6.3.8. Major Adverse Cardiovascular Events

MACE reported by investigators are adjudicated by an independent clinical endpoint committee in a blinded fashion. Unreported events may also be independently identified by the clinical endpoint committee.

The following positively adjudicated MACE will be considered as AESIs:

- death due to cardiovascular AEs
- myocardial infarction
- hospitalization for unstable angina
- hospitalization for heart failure
- coronary interventions (such as coronary artery bypass graft or percutaneous coronary intervention), and

- cerebrovascular events, including cerebrovascular accident (stroke) and transient ischemic attack.

The counts and percentages of participants with adjudicated MACE may be summarized by treatment. In addition, MACE reported by investigator may also be summarized although a MACE reported by investigator that is not positively adjudicated is not considered an AESI.

A listing of participants reporting MACE events, either reported by investigator or identified by the clinical endpoint committee, may be provided.

4.6.3.9. Major Depressive Disorder/Suicidal Ideation or Behavior

The severe/serious treatment-emergent major depressive disorder/suicidal ideation or behavior will be captured as AESI. AEs will be searched using MedDRA PT terms. Detailed searching criteria can be found in Appendix 2 (Section 6.2).

The counts and percentages of participants with TEAEs will be summarized by treatment group using MedDRA PT nested within SMQ. Events will be ordered by decreasing frequency in the total tirzepatide group nested within SMQ. A listing of participants with major depressive disorder/suicidal ideation or behavior may be provided if deemed necessary.

Additionally, suicidal ideation and behavior, and depression will be assessed by the investigator via spontaneously reported AEs and through the use of the C-SSRS and the PHQ-9.

4.6.3.9.1. Patient Health Questionnaire

Total scores for the PHQ-9 range from 0 to 27 with total scores categorized as

- none (not depressed): 0 through 4
- mild: 5 through 9
- moderate: 10 through 14
- moderately severe: 15 through 19, and
- severe: 20 through 27.

Shift tables will be provided showing the counts and percentages of participants within each baseline category (maximum value) versus each postbaseline category (maximum value) by treatment.

Additionally, the following 3 outcomes of interest will be compared between treatments (based on the maximum value during baseline and postbaseline):

- any increase in depression category (that is, worsening of depression): includes participants in the none, mild, moderate, or moderately severe category during baseline and with at least 1 postbaseline measurement
- increase from No or Mild Depression to Moderate, Moderately Severe, or Severe Depression: includes participants in the none or mild depression category during baseline and with at least 1 postbaseline measurement, and
- increase from Mild or Moderate Depression to Moderately Severe or Severe Depression: includes participants in the mild or moderate depression category during baseline and with at least 1 postbaseline measurement.

4.6.3.9.2. Suicidal Ideation and Behavior Solicited Through C-SSRS

Suicide-related thoughts and behaviors occurring during the entire study period, based on the C-SSRS, will be summarized by treatment group. In particular, for each of the following suicide-related events, the counts and percentages of participants with the event will be summarized by treatment group:

- died by suicide
- nonfatal suicide attempt
- interrupted attempt
- aborted attempt
- preparatory acts or behavior
- active suicidal ideation with specific plan and intent
- active suicidal ideation with some intent to act without specific plan
- active suicidal ideation with any methods (no plan) without intent to act
- nonspecific active suicidal thoughts
- wish to be dead, and
- nonsuicidal, self-injurious behavior.

In addition, the counts and percentages of participants who experienced at least 1 of the composite measures will be presented. The participants with at least 1 postbaseline C-SSRS assessment are included. The composite measure is determined at each assessment by the "yes" or "no" responses in C-SSRS categories by the study participant:

- Category 1 – Wish to be Dead
- Category 2 – Non-specific Active Suicidal Thoughts
- Category 3 – Active Suicidal Ideation with Any Methods (Not Plan) without Intent to Act
- Category 4 – Active Suicidal Ideation with Some Intent to Act, without Specific Plan
- Category 5 – Active Suicidal Ideation with Specific Plan and Intent
- Category 6 – Preparatory Acts or Behavior
- Category 7 – Aborted Attempt
- Category 8 – Interrupted Attempt
- Category 9 – Actual Attempt (non-fatal), and
- Category 10 – Completed Suicide.

Composite endpoints of suicidal ideation and suicidal behavior based on the above categories are defined below:

- **Suicidal ideation:** A "yes" answer at any time during study to any 1 of the 5 suicidal ideation questions (Categories 1 through 5) on the C-SSRS.
- **Suicidal behavior**: A "yes" answer at any time during study to any 1 of the 5 suicidal behavior questions (Categories 6 through 10) on the C-SSRS.
- **Suicidal ideation or behavior**: A "yes" answer at any time during study to any 1 of the 10 suicidal ideation and behavior questions (Categories 1 through 10) on the C-SSRS.

A listing contains data for each participant with suicidal ideation, suicidal behavior, or nonsuicidal self-injurious behavior during the study by treatment and visit. Data from all visits

are displayed, regardless of a "yes" or "no" answer, for participants with any "yes" answer for suicidal ideation, suicidal behavior, or self-injurious behavior without suicidal intent.

4.6.3.10. Renal Safety

Laboratory measures related to renal safety will be analyzed as specified for laboratory measurements in Section 4.6.5.

In addition, 2 shift tables examining renal function will be created. A minimum-to-mininum shift table of estimated glomerular filtration rate estimated by the Chronic Kidney Disease Epidemiology Collaboration equation with units mL/min/1.73 m^2, using categories (<30, ≥30 to <45, ≥45 to <60, ≥60 to <90, and ≥90 mL/min/1.73 m^2). A maximum-to-maximum shift table of UACR, using the categories UACR <30 mg/g, ≥30 mg/g UACR to ≤300 mg/g, UACR >300 mg/g (respectively, these represent normal, microalbuminuria, and macroalbuminuria).

MMRM analyses as described in Section 4.6 for estimated glomerular filtration rate and log-transformed UACR will be provided. Log transformation will be performed for UACR.

4.6.3.10.1. Acute Renal Events

Because severe GI events may lead to dehydration, which could cause a deterioration in renal function including acute renal failure, dehydration events will be analyzed as described in the next section. Acute renal events associated with chronic renal failure exacerbation will also be captured.

Severe/serious renal events from the following SMQ search will be considered as AESI.

The counts and percentages of participants with acute renal events may be summarized by treatment if overall count >10 by using the MedDRA PTs contained in any of the following SMQs:

- Acute renal failure: narrow terms in *Acute renal failure* SMQ (20000003), and
- Chronic kidney disease: narrow terms in *Chronic kidney disease* SMQ (20000213).

In addition, a listing of participants with treatment-emergent acute renal events may be provided, if deemed necessary.

4.6.3.10.2. Dehydration

Dehydration events will be captured in the narrow terms in *Dehydration* SMQ (20000232). Severe/serious dehydration events will be considered as AESI. A listing of participants with treatment-emergent dehydration events may be provided.

4.6.3.11. Thyroid Safety Monitoring

4.6.3.11.1. Calcitonin

The purpose of calcitonin measurements is to assess the potential of tirzepatide to affect thyroid C-cell function, which may indicate development of C-cell hyperplasia and neoplasms.

Observed calcitonin data (a thyroid-specific laboratory assessment) will be summarized by treatment and nominal visit.

The counts and percentages of participants with a maximum postbaseline calcitonin value in the following thresholds will be provided by treatment and maximum baseline calcitonin value (≤20 ng/L, >20 ng/L to ≤35 ng/L, >35 ng/L). Postbaseline categories are:

- ≤20 ng/L
- >20 ng/L to ≤35 ng/L
- >35 ng/L to ≤50 ng/L
- >50 ng/L to ≤100 ng/L, and
- >100 ng/L.

4.6.3.11.2. C-Cell Hyperplasia and Thyroid Malignancies

Treatment-emergent thyroid malignancies and C-cell hyperplasia will be considered as AESI. Thyroid malignancies and C-cell hyperplasia will be identified using MedDRA HLT for *Thyroid neoplasms* and PT for *Thyroid C-cell hyperplasia*.

The counts and percentages of participants with treatment-emergent thyroid C-cell hyperplasia and malignancies may be summarized or a listing of participants with treatment-emergent thyroid C-cell hyperplasia and neoplasms may be provided if deemed necessary.

4.6.3.12. Treatment-Emergent Arrhythmias and Cardiac Conduction Disorders

Severe/serious treatment-emergent arrhythmias and cardiac conduction disorders will be considered as AESI. The cardiovascular events will include clinically relevant rhythm and conduction disorders.

The treatment-emergent arrhythmias and cardiac conduction disorder events will be identified using the MedDRA PTs. Detailed searching criteria can be found in Appendix 2 (Section 6.2).

The counts and percentages of participants with treatment-emergent arrhythmias and cardiac conduction disorders may be summarized by treatment and PT nested within SMQ. The PT will be ordered with decreasing frequency in tirzepatide arm within SMQ. A listing of participants with treatment-emergent arrhythmias and cardiac conduction disorders may be provided if deemed necessary.

4.6.3.13. Treatment-Emergent Malignancy

The malignancy events will be included using the MedDRA PTs contained in the *Malignant tumours* SMQ (20000194) narrow scope or *Tumours of unspecified malignancy* SMQ (20000195) narrow scope. Malignancy will be considered as an AESI.

The counts and percentages of participants with treatment-emergent malignancy will be summarized by treatment.

4.6.3.14. Abuse Liability

To identify AE terms suggestive of abuse liability potential, narrow terms from the *Drug abuse and dependence* SMQ (20000101) will be used. The counts and percentages of participants will be summarized by treatment group in order of decreasing frequency.

These analyses will be performed for individual CSRs and the summary of clinical safety.

4.6.4. Vital Signs

In the case that multiple records of an individual vital sign are collected at the same visit, they will be averaged prior to being used for data summaries and analyses.

Descriptive summaries by treatment and by nominal visit will be provided for baseline and postbaseline values as well as change from baseline values.

An MMRM and/or an ANCOVA model, as discussed in Section 4.6, using data including from the safety follow-up period might be conducted if necessary.

Counts and percentages of participants with treatment-emergent abnormal sitting SBP, sitting diastolic blood pressure, and pulse at any time during the entire study (including the off-drug follow up time period) will be summarized by treatment for participants who have both baseline and at least 1 postbaseline result. A treatment-emergent high result is defined as a change from a value less than or equal to the high limit at baseline to a value greater than the high limit at any time that meets the specified change criteria during the postbaseline period. A treatment-emergent low result is defined as a change from a value greater than or equal to the low limit at baseline to a value less than the low limit at any time that meets the specified change criteria during the postbaseline period. To assess decreases, change from the minimum value during the baseline period to the minimum value during the postbaseline period will be used. To assess increases, changes from the maximum value during the baseline period to the maximum value during the postbaseline period will be used. Both planned and unplanned measurements will be included in the analysis. The criteria for identifying participants with treatment-emergent vital sign abnormalities are stated in Table GPIF.4.10.

Table GPIF.4.10. Categorical Criteria for Abnormal Treatment-Emergent Blood Pressure and Pulse Measurements

Parameter	Low	High
Systolic BP (mm Hg) (supine or sitting – forearm at heart level)	≤90 and decrease from baseline ≥20	≥140 and increase from baseline ≥20
Diastolic BP (mm Hg) (supine or sitting – forearm at heart level)	≤50 and decrease from baseline ≥10	≥90 and increase from baseline ≥10
Pulse (bpm) (supine or sitting)	<50 and decrease from baseline ≥15	>100 and increase from baseline ≥15

Abbreviations: BP = blood pressure; bpm = beats per minute

4.6.5. Clinical Laboratory Evaluation

Limits from the performing laboratory will be used to define low (L) and high (H). Descriptive summaries by treatment and by nominal visit will be provided for the baseline and postbaseline values as well as the change from baseline values. The associated descriptive will be presented in International System of Units and in conventional units.

For selected laboratory analyte measurements collected quantitatively, observed, and change from baseline values for each visit may be displayed in plots for participants who have both a baseline and at least 1 postbaseline planned measurement. Unplanned postbaseline measurements will be excluded from plots.

A shift table will be provided including unplanned measurements. The shift table will include the number and percentage of participants within each baseline category (low, normal, high, or missing) versus each postbaseline category (low, normal, high, or missing) by treatment. The proportion of participants shifted will be compared between treatments using Fisher's exact test.

For qualitative laboratory analytes, the number and percentage of participants with normal and abnormal values will be summarized by treatment.

A listing of abnormal findings will be created for laboratory analyte measurements, including qualitative measures. The listing will include:

- participant identification
- treatment group
- laboratory collection date
- study day
- analyte name, and
- analyte finding.

The MMRM model or ANCOVA (if MMRM model is not applicable) as described in Section 4.6 will be used for the analysis during the treatment period (excluding the safety follow-up period) for the continuous measurements for selected laboratory tests with or without log-transformed (postbaseline measure/baseline measure) response variables. For measures analyzed using log-transformed values, the results will be presented with the scale back transforming to the original, related scale.

The summary of treatment-emergent abnormal, high, or low laboratory results any time will be provided.

4.6.6. Product Complaints

A listing of all product complaints by subcategory, inclusive of device product complaints that led to an AE or that could have led to an SAE had intervention not been taken, will be provided.

4.7. Other Analyses

4.7.1. PAP Adherence

For participants in ISA2 only, the adherence to use of the PAP machine over the course of the study will be summarized. Specifically, the summary statistics for PAP adherence at baseline and at each postbaseline week, stratified by treatment arm will be provided. Additionally, the categorical shift in PAP adherence between baseline and Week 52, stratified by treatment arm using a shift table for increased, decreased, or stable PAP use will be assessed. Finally, the number and percentage of participants in ISA2 who withdraw from regular PAP use will be summarized.

4.7.2. Health Outcomes

The PRO questionnaires will be analyzed using the mITT population on the EAS, unless specified otherwise.

Item-level missingness will be dealt with per the instrument developers' instruction.

Additional psychometric analyses will be performed by Value, Evidence, and Outcomes at Lilly and documented in a separate analysis plan.

Analyses of actual and change from baseline in PRO scores will be conducted using linear models with baseline PRO scores, treatment, stratification factors, and other factors that may be considered relevant.

If an administrative error occurs where more than 1 PRO questionnaire is completed within the same visit window by the same participant with different responses on at least 1 item, the questionnaire with the worst response will be used (for example, the questionnaire with the highest PHQ-9 score will be used). If more than 1 PRO questionnaire is completed within the same visit window with the same response to each item, the most recent response will be used.

4.7.2.1. Functional Outcomes of Sleep Questionnaire

The FOSQ will be included to assess change in FOSQ domains and total score from baseline to Week 52. The FOSQ is a 30-item sleep-specific, participant-completed questionnaire used to assess the effect of disorders associated with excessive daytime sleepiness on daily functioning in adults. It assesses the following 5 domains of:

- General productivity (8 items)
- Activity level (9 items)
- Vigilance (7 items)
- Social outcomes (2 items), and
- Intimate and sexual relationships (4 items).

The FOSQ items assess participant's current status with each item rated on a scale of 1 (extreme difficulty) to 4 (no difficulty), with an additional not applicable (0 = "I don't do this activity for other reasons") also available. Individual domain scores are calculated by taking the mean of answered, non-zero items within each domain and ranges from 1 to 4 with higher scores indicating better outcomes. A total score can be calculated by first computing the mean score for each domain, then multiplying the mean of the domain scores by 5 (Weaver et al. 1997). The change from baseline in all 5 FOSQ domain scores will be assessed.

4.7.2.2. Functional Outcomes of Sleep Questionnaire, 10 Items

The FOSQ-10 will be included to assess change in FOSQ total score domains from baseline to Week 52. The 10-item sleep-specific, participant-completed questionnaire is a shortened version of the FOSQ with the same number of domains as the parent version. Of note, the FOSQ-10 has the same 5 domains as the FOSQ, but with fewer items per domain.

Calculation of the individual domain scores and the total score for the FOSQ-10 is carried out in a similar manner to FOSQ. The domain scores are first calculated by taking the mean of the answered, non-zero items within each domain. The total score is calculated by multiplying the mean of the domain scores by 5 (for each domain which has at least 1 response).

4.7.2.3. Epworth Sleepiness Scale

The ESS is an 8-item participant-completed measure that asks the participant to rate on a scale of 0 (would never doze) to 3 (high chance of dozing), their usual chances of dozing in 8 different daytime situations, with a recall period of "in recent times." The ESS total score is the sum of the 8-item scores and ranges from 0 to 24, with higher scores indicating greater daytime sleepiness (Johns 1991). Of note, if 1 or more item scores are missing, the ESS total score is not valid and will not be included in the analysis.

4.7.2.4. PROMIS Short Form v1.0 Sleep-Related Impairment 8a

The PROMIS Short Form v1.0 Sleep-Related Impairment 8a consists of 8 items each rated on a 5-point scale ranging from "not at all" to "very much." Items have a recall period of "in the past 7 days." Individual item scores will be combined and converted to a T-score using a response pattern scoring approach (Northwestern 2016a). The T-score standardizes the raw score to a distribution with a mean of 50 and standard deviation of 10.

4.7.2.5. PROMIS Short Form v1.0 Sleep Disturbance 8b

The PROMIS Short Form v1.0 Sleep Disturbance 8b consists of 8 items each rated on a 5-point scale ranging from "not at all" to "very much," "never" to "always," or "very poor" to "very good." Items have a recall period of "in the past 7 days." Individual item scores will be combined and converted to a T-score using a response pattern scoring approach (Northwestern 2016b). The T-score standardizes the raw score to a distribution with a mean of 50 and standard deviation of 10.

For item 8 of this scale (which is a measure of sleep quality), counts and percentages of participants at each time point will be summarized by nominal visit and by treatment. A shift table from baseline to postbaseline will be created at each postbaseline visit.

4.7.2.6. Short-Form-36 Health Survey Version 2, Acute Form

Per copyright owner, the QualityMetric Health Outcomes™ Scoring (PRO_CoRe V2.0) Software will be used to derive the following domain and component scores:

- Mental Component Score (MCS)
- Physical Component Score (PCS)
- Physical Functioning domain (PF)
- Role-Physical domain (RP)
- Bodily Pain domain (BP)
- General Health domain (GH)
- Vitality domain (VT)
- Social Functioning domain (SF)
- Role-Emotional domain (RE), and
- Mental Health domain (MH).

The following analyses for the actual value and change from baseline value for each domain and component score will be conducted:

- descriptive summaries by treatment group, and
- analysis described in Table GPIF.4.3.

4.7.2.7. Patient Global Impression of Status/Change for OSA Outcomes

The counts and percentages of participants for PGIS for Physical Activity and PGIC response categories at each time point will be summarized by nominal visit and by treatment. A shift table from baseline to postbaseline of 3 PGIS response categories (OSA Sleepiness, Fatigue, and Snoring) and 4 PGIC response categories (OSA Sleepiness, Fatigue, Snoring, and Sleep Quality) will be created at each postbaseline visit.

4.7.2.8. EQ-5D-5L

For the utility score and the Visual Analog Scale scores, the following analyses of the actual value and change from baseline value will conducted:

- descriptive summaries by treatment group, and
- ANCOVA described in Table GPIF.4.4.

4.7.3. Subgroup Analyses

The following subgroups will be analyzed using the efficacy estimand on change in AHI values from baseline to 52-week visit if there is an adequate number of participants in each treatment by subgroup (for example, 10%):

- age (<50 years, ≥50 years)
- baseline OSA severity (not severe, severe)
- race
- ethnicity
- region of enrollment (US, OUS)
- gender (male or female)
- baseline BMI (<35, ≥35 and <40, ≥40 kg/m^2), and
- baseline ESS (ESS ≤10, ESS >10).

Analyses for change from baseline in AHI will be performed using an MMRM model that includes the same fixed effects given for the primary analysis model plus factors of subgroup, 2-way interaction of subgroup and treatment, 2-way interaction of subgroup and visit, and 3-way interaction of treatment, visit, and subgroup. The possible interaction effect of subgroup and treatment at the primary endpoint (Week 52) will be evaluated. When analyzing OSA severity (not severe, severe) as a subgroup, the baseline AHI will not be included as a covariate to avoid confounding.

4.8. Interim Analyses

The details for the interim analyses and Data Monitoring Committee will be provided in the Data Monitoring Committee Charter.

4.8.1. Unblinding Plan

Details of the blinding and unblinding are provided in the Blinding and Unblinding Plan document for Master Protocol GPIF.

4.9. Changes to Protocol-Planned Analyses

Refer to Table GPIF.1.1.

5. Sample Size Determination

Approximately 206 participants per ISA will be randomly assigned to either tirzepatide or placebo in a 1:1 ratio (approximately 103 participants per treatment arm), and the statistical power will be evaluated for the primary efficacy endpoint and key secondary combination PRO endpoint at a 2-sided significance level of 0.05. This sample size will provide the following:

- at least 90% power to demonstrate superiority of tirzepatide at the MTD (10 mg or 15 mg) to placebo for the mean percent change from baseline in AHI, assuming 50% improvement, with a common standard deviation of 50% and a dropout rate of 25%, and
- at least 90% power to demonstrate superiority of tirzepatide at the MTD (10 mg or 15 mg) to placebo for the hierarchical combination PRO endpoint using the Finkelstein-Schoenfeld method (Finkelstein and Schoenfeld 1999), with a dropout rate of 25%.

An upper limit of approximately 70% enrollment of male participants will be used to ensure a sufficiently large sample of female participants.

6. Supporting Documentation

6.1. Appendix 1: Clinical Trial Registry Analyses

Additional analyses will be performed for the purpose of fulfilling the Clinical Trial Registry requirements.

Analyses provided for the Clinical Trial Registry requirements include the following:

- Summary of AEs, provided as a dataset which will be converted to an XML file. Both SAEs and 'Other' non-SAEs are summarized: by treatment group, by MedDRA PT.
    - An AE is considered 'Serious' whether or not it is a TEAE.
    - An AE is considered in the 'Other' category if it is both a TEAE and is not serious. For each SAE and 'Other' AE, for each term and treatment group, the following are provided:
        - the number of participants at risk of an event
        - the number of participants who experienced each event term, and
        - the number of events experienced.
- For each SAE, these additional terms are provided for EudraCT:
    - the total number of occurrences causally related to treatment
    - the total number of deaths, and
    - the total number of deaths causally related to treatment.
- Consistent with www.ClinicalTrials.gov requirements, 'Other' AEs that occur in fewer than 5% of participants/subjects in every treatment group may be excluded.
- AE reporting is consistent with other document disclosures for example, the CSR, manuscripts, and so forth.

Demographic table including the following age ranges required by EudraCT: adults (18 to 64 years), (65 to 85 years), and (85 years and over).

6.2. Appendix 2: Search Criteria for Special Safety Topics

Arrhythmias and cardiac conduction disorders

Treatment-emergent arrhythmias, arrhythmias and cardiac conduction disorders will be considered an AESI. The cardiovascular events will include clinically relevant rhythm and conduction disorders. The treatment-emergent arrhythmias and cardiac conduction disorders events will be included using the MedDRA PT contained in any of the following SMQs:

1) Arrhythmias:
    a. For symptoms: *Arrhythmia related investigations, signs and symptoms* SMQ (20000051), narrow and broad terms

Approved on 28 Feb 2024 GMT

b. For supraventricular arrhythmias: In *Cardiac arrhythmia* SMQ, under tachyarrhythmia sub SMQ
    i. *Supraventricular tachyarrhythmia* SMQ (20000057), broad and narrow terms
    ii. *Tachyarrhythmia terms, nonspecific* SMQ (20000164), narrow terms only; and
    iii. *Ventricular tachyarrhythmia* SMQ (20000058), narrow terms only.

2) Cardiac conduction disorders
    a. *Conduction defects* SMQ (20000056), narrow terms only; and
    b. *Cardiac conduction disorders* HLT (10000032), all PTs.

Hepatic TEAEs

Treatment-emergent potentially drug-related hepatic disorders will be summarized by treatment using the MedDRA PTs contained in any of the following SMQs:

- broad and narrow terms in the *Liver related investigations, signs and symptoms* SMQ (20000008)
- broad and narrow terms in the *Cholestasis and jaundice of hepatic origin* SMQ (20000009)
- broad and narrow terms in the *Hepatitis non-infections* SMQ (20000010)
- broad and narrow terms in the *Hepatic failure, fibrosis and cirrhosis and other liver damage* SMQ (20000013), and
- narrow terms in the *Liver-related coagulation and bleeding disturbances* SMQ (20000015).

Acute gallbladder disease

All events of TEAE biliary colic, cholecystitis, or other suspected events related to gallbladder disease will be summarized by treatment groups by PT with decreasing frequency under the following SMQs:

- narrow PTs in *Gallbladder related disorders* SMQ (20000124)
- narrow PTs in *Biliary tract disorders* SMQ (20000125), and
- narrow PTs in *Gallstone related disorders* SMQ (20000127).

Major depressive disorder/suicidal ideation

AEs will be searched using MedDRA PTs from SMQs narrow scope: 20000037 (*Suicide/self-injury*) and 20000167 (*Depression [excl suicide and self injury]*).

C-cell hyperplasia and thyroid malignancies

Thyroid malignancies and C-cell hyperplasia will be identified using MedDRA HLT for *Thyroid neoplasms* and PT for *Thyroid C-cell hyperplasia*.

Hypersensitivity reactions

Analyses are based on the following:

- narrow and algorithm terms in *Anaphylactic reaction* SMQ (20000021)
- narrow terms in *Angioedema* SMQ (20000024)
- narrow terms in *Severe cutaneous adverse reactions* SMQ (20000020), and
- narrow terms in *Hypersensitivity* SMQ (20000214).

For the Anaphylactic reaction SMQ, each term is classified by scope (narrow, broad) and by category (A, B, C, and D). All narrow terms are category A, and all broad terms are category B, C, or D. In addition to the usual narrow and broad searches, the SMQ defines an algorithm to further refine the cases of interest. For Time Period A analysis, the *Anaphylactic reaction* SMQ algorithm will be included. The algorithm is based upon events that occur within Time Period A. The counts and percentages of participants who experienced a TEAE for the following will be analyzed for each of the 2 time periods:

- any narrow term from any 1 of the 4 SMQs indicated above (that is, combined search across narrow of all 4 SMQs), and
- any narrow scope term within each SMQ, separately (that is, narrow SMQ search). For Time Period A analysis, any term from *Anaphylactic reaction* SMQ algorithm.

Injection site reactions

The ISR AE will be identified using the MedDRA PT in any of the following:

- HLT of *Injection site reaction*
- HLT of *Administration site reaction*, and
- HLT of *Infusion site reactions*.

Pancreatitis events

Determination of investigator-reported events will be through the *Acute pancreatitis* MedDRA SMQ (20000022, narrow scope) and a *Chronic pancreatitis* PT search of the AE database, while adjudication-confirmed pancreatitis is found from adjudication forms.

6.3. Appendix 3: Magnetic Resonance Imaging Addendum

This section is applicable to the participants who are enrolled in the MRI addendum.

This addendum applies to a subset of participants (approximately 58 participants) enrolled in ISA1. MRIs for the assessment of fat dispositions will be collected at baseline and Week 52. The MRI at baseline needs to be completed prior to Visit 2 or within 7 days after Visit 2. The MRI at Week 52 may be scheduled for any day ± 14 days.

MRI analyses will be guided by the treatment policy strategy and conducted among all participants who are enrolled in the addendum, received at least 1 dose of study drug, and have baseline and at least 1 postbaseline MRI measurement. No imputation will be performed for

missing data. The participant's demographics and baseline characteristics for the MRI addendum will be summarized.

Objectives	Endpoints	Analytical Approaches
Compare the effect of once weekly tirzepatide at MTD versus placebo on the changes of soft tissues volumes, fat volumes and fat content (%) in upper airway structures and in the abdomen in participants with OSA and obesity.	Changes of absolute soft tissue volume, fat volume and fat content (%) of the following: • Tongue • Soft palate • Pterygoid muscle • Lateral pharyngeal walls • Neck and submandibular area • Total, visceral and subcutaneous abdominal fat	Change from baseline to Week 52 for each parameter will be compared between treatment arms using an ANCOVA approach. The model will include treatment, the stratification factors of gender, and baseline AHI (not severe/severe), and the baseline value for the parameter. Summary statistics for MRI parameters at baseline and at Week 52 will be provided. The treatment comparison at baseline will be performed using an ANOVA model.
Explore correlation of changes of soft tissue volumes, fat volumes, and fat content (%) in upper airway structures and in the abdomen with changes of AHI.	Correlations between the change in absolute soft tissue volume, fat volume, and fat content (%) for the structures listed above and the % change in AHI.	Spearman correlations between the change from baseline for each of the MRI endpoints and the % change in AHI will be calculated.

Abbreviations: AHI = Apnea-Hypopnea Index; ANCOVA = analysis of covariance; ANOVA = analysis of variance; MRI = magnetic resonance imaging; MTD = maximum tolerated dose; OSA = obstructive sleep apnea.

6.4. Appendix 4: Statistical Analysis for China

Analyses will be performed for the following subpopulations:

- participants enrolled in East Asian countries/regions (China, Japan), and
- participants enrolled in China.

The analysis methods for the above-mentioned subgroups will be similar to those described for the main part of this SAP. If there is not a sufficient number of participants in the subpopulation, summary statistics will be provided.

The analyses to be included will be documented in a separate list of analyses which should include disposition, demographics, and selected efficacy and safety endpoints.

6.5. Appendix 5: Statistical Analysis for Japan

Analyses will be performed for the following subpopulations:

- participants enrolled in Japan, and
- the JASSO subpopulation, participants who meet the criteria of the JASSO (not limited to the participants enrolled in Japan).

The JASSO subpopulation analysis will be performed according to the criteria of both BMI and obesity-related health problems according to the treatment flow of obesity disease in the obesity disease treatment guideline (JASSO 2022). The JASSO guideline states that drug treatment in

participants with obesity disease should follow with a BMI above or equal to 25 kg/m^2 (in this trial, BMI enrollment was started from 27 kg/m^2), and at least 2 obesity-related health problems, or a BMI above or equal to 35 kg/m^2 and at least 1 obesity-related health problem out of the 11 obesity-related health problems, including OSAS, listed below. The overall population and participants with obesity disease according to the JASSO guideline will be compared.

The analysis methods for the above-mentioned subgroups will be similar to those described for the main part of this SAP. If there is not a sufficient number of participants in the subpopulation, summary statistics will be provided.

As a low number of participants were enrolled from Japan, combined analyses with both ISAs may be conducted to explore a future line extension for the OSAS indication in Japan.

The analyses to be included will be documented in a separate list of analyses which should include disposition, demographics, and selected efficacy and safety endpoints.

Eleven obesity-related health problems

The JASSO guideline defines 11 health problems for the diagnosis of "Obesity Disease" in subjects who need weight reduction for a medical reason. Data collected by a specific CRF will be used to specify the obesity-related health problems for each participant. The list of the 11 health problems are:

1) Glucose intolerance disorder (type 2 diabetes, impaired glucose tolerance [IGT], and so on)
2) Dyslipidemia
3) Hypertension
4) Hyperuricemia and Gout
5) Cardiovascular disease, myocardial infarction and Angina
6) Cerebral infarction and transient ischemic attack (TIA)
7) Non-alcoholic fatty liver disease (NAFLD)
8) Menstruation disorder and infertility
9) Obstructive sleep apnea syndrome (OSAS) and obesity-hypoventilation syndrome
10) Motor dysfunction: arthritis/osteoarthritis (knee, hip joint, supine, and so on), and
11) Obesity-related renal disease.

7. References

Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med*. 2014;33(4):693-713. https://doi.org/10.1002/sim.5974

Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med*. 2009;28(4):586-604. https://doi.org/10.1002/sim.3495

Bretz F, Posch M, Glimm E, et al. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biom J*. 2011;53(6):894-913. https://doi.org/10.1002/bimj.201000239

Bretz F, Xi D. Commentary on "Statistics at FDA: reflections on the past six years". *Stat Biopharm Res*. 2019;11(1):20-25. https://doi.org/10.1080/19466315.2018.1562369

[ADA] American Diabetes Association. Glycemic targets: standards of medical care in diabetes—2020. *Diabetes Care*. 2020:43(suppl 1):S66–S76. https://doi.org/10.2337/dc20-S006

Danne T, Philotheou A, Goldman D, et al. A randomized trial comparing the rate of hypoglycemia – assessed using continuous glucose monitoring – in 125 preschool children with type 1 diabetes treated with insulin glargine or NPH insulin (the PRESCHOOL study). *Pediatr Diabetes*. 2013;14(8):593-601. https://doi.org/10.1111/pedi.12051

Donovan LM, Yu L, Bertisch SM, et al. Responsiveness of patient-reported outcomes to treatment among patients with type 2 diabetes mellitus and OSA. *Chest.* 2020;157(3):665-672. https://doi.org/10.1016/j.chest.2019.11.011

Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Stat Med.* 1999;18(11):1341-1354. https://doi.org/10.1002/(sici)1097-0258(19990615)18:11<1341::aid-sim129>3.0.co;2-7

[JASSO] Japan Society for the Study of Obesity. Guideline for the management of obesity disease. Tokyo: Life Science Publishing. 2022. (in Japanese)

Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*. 1991;14(6):540-545. https://doi.org/10.1093/sleep/14.6.540

[Northwestern]. Northwestern University. PROMIS Short Form v1.0 – Sleep-Related Impairment 8a. Published May 09, 2016a. Accessed December 14, 2022. https://www.healthmeasures.net/index.php?option=com_instruments&view=measure&id=185&Itemid=992

[Northwestern]. Northwestern University. PROMIS Short Form v1.0 – Sleep Disturbance 8b. Published May 03, 2016b. Accessed December 14, 2022. https://www.healthmeasures.net/index.php?option=com_instruments&view=measure&id=184&Itemid=992

Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33(2):176-182. https://doi.org/10.1093/eurheartj/her352

Rubin DB. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons Inc.; 1987.

Vandemeulebroecke M, Häring DA, Hua E, et al. New strategies for confirmatory testing of secondary hypotheses on combined data from multiple trials [published online February 4, 2024]. *Clin Trials*. https://doi.org/10.1177/17407745231214382

Weaver TE, Laizner AM, Evans LK, et al. An instrument to measure functional status outcomes for disorders of excessive sleepiness. *Sleep*. 1997;20(10):835-843. https://doi.org/10.1093/SLEEP%2F20.10.835

Weinberg ME, Bacchetti P, Rushakoff RJ. Frequently repeated glucose measurements overestimate the incidence of inpatient hypoglycemia and severe hyperglycemia. *J Diabetes Sci Technol*. 2010;4(3):577-582. https://doi.org/10.1177/193229681000400311

Ye T, Bannick M, Yi Y, Shao J. Robust variance estimation for covariate-adjusted unconditional treatment effect in randomized clinical trials with binary outcomes. *Stat Theory Relat Fields*. 2023;7(2):159-163. https://doi.org/10.1080/24754269.2023.2205802

Signature Page for VV-CLIN-117998 v2.0

Approval	PPD

Signature Page for VV-CLIN-117998 v2.0