

Official Title: A Phase III, Randomized, Double-Masked, Placebo-Controlled, Multicenter Study to Evaluate the Efficacy, Safety, Pharmacokinetics, and Pharmacodynamics of Satralizumab in Participants with Moderate-to-Severe Thyroid Eye Disease

NCT Number: NCT06106828

Document Date: Statistical Analysis Plan Version 3: 25-Jul-2025

STATISTICAL ANALYSIS PLAN

STUDY TITLE: PHASE III, RANDOMIZED, DOUBLE-MASKED, PLACEBO-CONTROLLED, MULTICENTER STUDIES TO EVALUATE THE EFFICACY, SAFETY, PHARMACOKINETICS, AND PHARMACODYNAMICS OF SATRALIZUMAB IN PARTICIPANTS WITH MODERATE-TO-SEVERE THYROID EYE DISEASE

STUDY NUMBER: GP44467 and GP44729

STUDY NAME: SatraGO-1 and SatraGO-2

VERSION NUMBER: 3

ROCHE COMPOUND: Satralizumab (RO5333787)

EUDRACT NUMBER: 2023-503309-13-00 and 2023-503669-50-00

IND NUMBER: 163819

NCT NUMBER: NCT05987423 and NCT06106828

PLAN PREPARED BY: [REDACTED]

STATISTICAL ANALYSIS PLAN APPROVAL

SPONSOR: F. Hoffmann-La Roche Ltd

LEGAL REGISTERED ADDRESS: Grenzacherstrasse 124
4070 Basel, Switzerland

DATE FINAL: See electronic date stamp on the last page of this document

CONFIDENTIAL

This is an F. Hoffmann-La Roche Ltd document that contains confidential information. Nothing herein is to be disclosed without written consent from F. Hoffmann-La Roche Ltd

STATISTICAL ANALYSIS PLAN VERSION HISTORY

This SAP was developed based on Roche SAP eMD_v2.0.

SAP Version	Approval Date	Based on Protocol (Version, Approval Date)
1	15 April 2025	GP44467: V3.0 GP44729: V3.1
2	11 June 2025	GP44467: V3.0 GP44729: V3.1
3	Refer to electronic signature and date stamp on the final page of this document	GP44467: V3.0 GP44729: V3.1

STATISTICAL ANALYSIS PLAN AMENDMENT RATIONALE

Key changes to the SAP V3, along with the rationales for each change, are summarized below.

Section	Description of Change	Rationale for Change

Additional minor changes have been made throughout to improve clarity and consistency.

TABLE OF CONTENTS

STATISTICAL ANALYSIS PLAN AMENDMENT RATIONALE.....	3
1. INTRODUCTION.....	9
1.1 Study Rationale	9
1.2 Background on Thyroid Eye Disease.....	9
1.3 Background on Satralizumab.....	10
1.4 Objectives, Estimands and Endpoints	10
1.5 Study Design	20
1.5.1 Overview of Study Design.....	21
1.5.2 Part I Period.....	21
1.5.3 Part II Period.....	21
1.5.4 Follow-up Period.....	22
1.5.5 Treatment Assignment.....	23
1.5.6 Data Monitoring	23
2. ANALYSIS SETS	24
3. STATISTICAL HYPOTHESES AND SAMPLE SIZE DETERMINATION	25
3.1 Statistical Hypotheses	25
3.1.1 Analyses for the European Medicines Agency.....	26
3.1.2 Analyses for the Japan Pharmaceuticals and Medical Devices Agency	27
3.2 Sample Size Determination	28
4. STATISTICAL ANALYSES	29
4.1 General Considerations	29
4.2 Primary Estimand Analysis	31
4.2.1 Definition of Primary Endpoint/Estimand	31
4.2.2 Main Analytical Approach for the Primary Endpoint.....	32
4.2.3 Sensitivity Analyses	34
4.2.4 Supplementary Analyses	34
4.3 Stratification Factors	35
4.4 Secondary Endpoints.....	36
4.4.1 Key Secondary Endpoints.....	36

4.4.1.1	Handling of Missing Data for Key Secondary Endpoints:.....	37
4.4.2	Secondary Endpoints.....	37
4.4.2.1	Handling of Missing Data for Secondary Endpoints:.....	37
4.4.2.2	Binary endpoints	37
4.4.2.3	Continuous Endpoints:.....	39
4.4.2.4	Non-Comparative Analyses	39
4.5	Exploratory Endpoint Analysis	40
4.5.1.1	Handling of Missing Data for Exploratory Endpoints:.....	40
4.5.1.2	Binary Endpoints.....	40
4.5.1.3	Continuous Endpoints.....	42
4.5.1.4	Non-Comparative Analyses	42
4.6	Hypotheses testing for Part II.....	44
4.7	Safety Analyses	48
4.7.1	Extent of Exposure	49
4.7.2	Adverse Events.....	49
4.7.2.1	Definitions	50
4.7.2.2	Laboratory Data	52
4.7.2.3	Vital Signs.....	53
4.7.2.4	ECGs	53
4.8	Other Analyses	53
4.8.1	Summaries of Conduct of Study	53
4.8.2	Summaries of Demographics and Baseline Characteristics	53
4.8.3	Pharmacokinetic Analyses.....	53
4.8.4	Immunogenicity Analyses	54
4.8.4.1	Pharmacodynamic Analyses.....	54
4.9	Interim Analyses	54
4.9.1	Safety Review	54
5.	SUPPORTING DOCUMENTATION	55
6.	REFERENCES.....	56

LIST OF TABLES

Table 1	Primary and Key Secondary Objectives and Corresponding Estimands	11
Table 2	Other Secondary Objectives and Endpoints	15
Table 3	Exploratory Objectives and Endpoints	18
Table 4	Participant Analysis Sets	24

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

Abbreviation or Term	Description
ADA	anti-drug antibody
AE	adverse event
AESI	adverse event of special interest
AQP-4	aquaporin-4
AUC	area under the concentration–time curve
CAS	clinical activity score
CI	Confidence interval
CS	Corticosteroid
CSR	Clinical Study Report
EUGOGO	European Group on Graves' Orbitopathy
FAS	Full analysis set
GO	Graves' ophthalmopathy
GO-QoL	Graves' Ophthalmopathy Quality of Life
ICE	Intercurrent event
ICH	International Council on Harmonization
IDCC	Independent Data Coordinating Center
iDMC	independent Data Monitoring Committee
IL-6	interleukin-6
IOP	Intraocular pressure
IxRS	interactive voice/web-based response system
IV	Intravenous
LSM	Least square means
MAR	Missing at random
MedDRA	Medical Dictionary for Regulatory Activities
MH	Mantel-Haenszel
MMRM	Mixed-model repeated measurements
NCI CTCAE	National Cancer Institute Common Terminology Criteria for Adverse Events
NMOSD	Neuromyelitis optica spectrum disorder
OSDI	ocular surface disease index;
PD	Pharmacodynamic
PK	Pharmacokinetic
PLB	Placebo
PR	Proptosis response
PT	Preferred term

Q4W	every four weeks
SAE	serious adverse events
SAP	Statistical Analysis Plan
SAS	Safety analysis set
SAT	Satralizumab
SD	Standard deviation
SC	Subcutaneous
SOC	System organ class
TEAE	Treatment-emergent adverse event
TED	thyroid eye disease
TSHR	thyrotropin (TSH) receptor
VA	Visual acuity
V/F	apparent volume of distribution.

1. INTRODUCTION

1.1 STUDY RATIONALE

The purpose of Studies GP44467 and GP44729 is to assess the efficacy, safety, pharmacokinetics and pharmacodynamics of subcutaneous (SC) satralizumab, a recombinant, humanized anti-interleukin (IL)-6 receptor (IL-6R) monoclonal antibody, in participants with moderate-to-severe active and chronic inactive thyroid eye disease (TED), a complex orbital inflammatory autoimmune disorder. In the United States and a few other countries, intravenous (IV) teprotumumab is approved for the treatment of TED but it is associated with a high rate of disease relapse or unfavorable side effects ([Allen et al. 2021](#)). In the rest of world, high-dose corticosteroids (CS) are the mainstay of treatment for active, moderate-to-severe TED, but are also associated with disease relapses and substantial side effects ([Patel et al. 2019](#)). Additionally, there is no established medical treatment for chronic inactive TED other than surgical intervention. Given that existing treatment options are limited, not disease modifying and associated with severe adverse events, there is an unmet medical need for new and disease-modifying therapies in patients with active moderate-to-severe TED and in those with chronic inactive TED.

The analyses specified in this document supersede the analysis plan described in the study protocols.

1.2 BACKGROUND ON THYROID EYE DISEASE

TED, also referred to as Graves' ophthalmopathy or orbitopathy (GO) or thyroid-associated orbitopathy, is a chronic inflammatory disorder predominantly associated with Graves' disease ([Bartley 1994](#)). While 90% of TED is associated with Graves' disease, 5% occurs in hypothyroid state and 5% in euthyroid individuals. The clinical presentation of TED is highly variable in severity, ranging from mild disease which may be self-limiting, to severe sight-threatening disease in 3%-7% of cases ([Wiersinga and Bartalena 2002](#); [Barrio-Barrio et al. 2015](#)). Moderate-to-severe disease can be associated with permanent visual impairment, as well as morphologic disfiguring changes to the face, related to proptosis, lid lag, and gaze disturbances, all which impact patient quality of life ([Gerding et al. 1997](#); [Kahaly et al. 2002](#); [Wiersinga 2012](#)).

TED is an autoimmune disease and although the underlying pathophysiological mechanism is not completely understood, it is presumed to be mediated through the activation of orbital fibroblasts by autoantibodies directed against thyroid antigens, such as thyrotropin receptor (TSH receptor; TSHR) autoantibodies and insulin-like growth factor-1 receptor autoantibodies ([Ezra et al. 2012](#)). Typically, TED has a progressive (active) inflammatory phase, which eventually leads to the emergence of fibrosis and a mixed inflammatory or fibrotic phase followed by a stable (inactive) post-inflammatory phase. In the early, active inflammatory stages of TED, which are reported to typically last from up to 1 year ([Bahn 2010](#)) to 3 years ([Douglas et al. 2022a](#)), studies have found both focal and diffuse inflammatory infiltration of the orbital tissues. Extraocular and

levator muscles, and adipose tissues are infiltrated by mononuclear cells such as CD4+ and CD8+ T lymphocytes, B cells and macrophages (Weetman et al. 1989; Gupta and Douglas 2011; Khong et al. 2016). T cells produce inflammatory cytokines such as IL-2 and IL-6, interferon- γ and the tumor necrosis factors, activating the orbital fibroblasts which proliferate and produce collagen and glycosaminoglycans responsible for the tissue expansion seen in TED (Dik et al. 2016). In addition, orbital fibroblasts differentiate into new mature fat cells further increasing orbital tissue volume. In a disease of longer duration, the production of inflammatory mediators expands, as type 2 helper T cells predominate and produce IL-5 and IL-10. Other inflammatory mediators such as IL-1, IL-6, IL-16, and transforming growth factor α [TGF- α] are produced with the involvement of macrophages, fibroblasts, and adipocytes in the orbit as well (Bahn 2012).

1.3 BACKGROUND ON SATRALIZUMAB

Satralizumab is a subcutaneously administered recycling humanized anti-IL-6R monoclonal antibody. Satralizumab shows pharmacological effects of IL-6R blockade at an every 4-week dosing frequency by SC injection, supported by its prolonged half-life compared with a typical IgG antibody, due to its recycling property. The drug achieves its pharmacological effects by binding to the IL-6R, thereby, inhibiting IL-6 signaling and the actions of IL-6.

Satralizumab is approved for the treatment of patients with neuromyelitis optica spectrum disorder (NMOSD) who are aquaporin-4 (AQP4) antibody positive. In a limited number of countries, satralizumab was approved for the treatment of patients with NMOSD, irrespective of their AQP4-IgG serostatus.

Detailed information on satralizumab is provided in the satralizumab Investigator's Brochure.

1.4 OBJECTIVES, ESTIMANDS AND ENDPOINTS

The studies (GP44467 and GP44729) will evaluate the efficacy, safety, pharmacokinetics and pharmacodynamics of satralizumab compared with placebo in participants with moderate-to-severe TED.

Table 1 presents the primary and key secondary objectives for the studies expressed using the estimand framework in accordance with the International Council for Harmonisation (ICH) E9(R1) statistical principles for clinical trials (ICH 2020). Table 2 and Table 3 presents the remaining objectives and corresponding endpoints.

Table 1 Primary and Key Secondary Objectives and Corresponding Estimands

Primary Objective	Estimand Definition
To evaluate the efficacy of subcutaneous satralizumab compared with placebo at Week 24	- Population: adult participants with moderate to severe active thyroid eye disease - Variable: proportion of participants achieving ≥ 2mm reduction in proptosis from baseline (Day 1) at Week 24 in the study eye, provided there is no deterioration of proptosis (≥ 2mm increase) in the fellow eye - Treatment (as randomized): - Experimental: satralizumab SC [REDACTED] mg, [REDACTED] mg, or [REDACTED] mg (for bodyweight [REDACTED] respectively), at Day 1, Week 2, and Week 4 (loading doses) and maintenance doses Q4W thereafter until Week 20. Seven doses in total during the 24-week Part I period - Control: placebo SC at Day 1, Week 2, and Week 4 (loading doses), and maintenance doses Q4W thereafter until Week 20. Seven doses in total during the 24-week Part I period - Intercurrent events occurring prior to Week 24 and handling strategies: - Early discontinuation of study treatment due to AEs or lack of efficacy: treatment policy strategy - Use of any prohibited systemic treatment or prohibited therapy for either eye (study eye or fellow eye): evaluated as non-responder (composite strategy) - Administration of rescue treatment: evaluated as non-responder (composite strategy) - Orbital or ocular surgery/procedure for TED (e.g., orbital decompression, eyelid or strabismus surgery, radiotherapy) in either eye (study eye or fellow eye): evaluated as non-responder (composite strategy) - Population-level summary: difference in proportions of proptosis responders at Week 24 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights

Table 1 Primary and Key Secondary Objectives and Corresponding Estimands (cont.)

Key Secondary Analyses Objectives	Estimand Definition
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on proptosis response	- Population: adult participants with moderate to severe active or chronic inactive TED - Variable: proportion of participants achieving ≥ 2 mm reduction in proptosis from baseline (Day 1) at Week 24 in the study eye provided there is no corresponding deterioration in proptosis (≥ 2 mm increase) in the fellow eye - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: as in primary estimand definition - Population-level summary: difference in proportions of proptosis responders in overall (active and chronic inactive) population at Week 24 between the satralizumab and placebo arm, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on change in proptosis	- Population: adult participants with moderate to severe active TED - Variable: Change from baseline (Day 1) in proptosis at Week 24 - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: intercurrent events as defined above. A treatment policy strategy will be applied to all intercurrent events ^a - Population-level summary: least square mean difference between the satralizumab and placebo group as evaluated using MMRM
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on change in proptosis	- Population: adult participants with moderate to severe active or chronic inactive TED - Variable: change from baseline (Day 1) in proptosis at Week 24 - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: intercurrent events as defined above. A treatment policy strategy will be applied to all intercurrent events ^a - Population-level summary: least square mean difference between the satralizumab and placebo group as evaluated using MMRM

Table 1 Primary and Key Secondary Objectives and Corresponding Estimands (cont.)

Key Secondary Objectives (cont.)	Estimand Definition
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on reduction/improvement in diplopia	- Population: adult participants with active TED with baseline (Day 1) diplopia > 0 - Variable: proportion of participants achieving ≥ 1 grade reduction/improvement in diplopia at Week 24 - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: as in primary estimand definition - Population-level summary: difference in proportions of participants achieving ≥ 1 grade reduction/improvement in diplopia at Week 24 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on reduction/improvement in diplopia	- Population: subset of overall (active and chronic inactive) TED participants with baseline (Day 1) diplopia > 0 - Variable: proportion of participants achieving ≥ 1 grade reduction/improvement in diplopia at Week 24 - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: as in primary estimand definition - Population-level summary: difference in proportions of participants achieving ≥ 1 grade reduction/improvement in diplopia at Week 24 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on improvement in orbital pain	- Population: subset of active TED participants with motility-induced pain at baseline (Day 1) in the study eye - Variable: proportion of participants achieving absence of motility-induced pain at Week 24 - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: as in primary estimand definition - Population-level summary: difference in proportions of participants with absence of motility-induced pain at Week 24 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights

Table 1 Primary and Key Secondary Objectives and Corresponding Estimands (cont.)

Key Secondary Objectives (cont.)	Estimand Definition
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on improvement in orbital pain	- Population: subset of active TED with spontaneous pain at baseline (Day 1) in the study eye - Variable: proportion of participants achieving absence of spontaneous pain at Week 24 - Treatment: as defined above - Intercurrent events occurring prior to Week 24 and handling strategies: as in primary estimand definition - Population-level summary: difference in proportions of participants with absence of spontaneous pain at Week 24 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights

AE = adverse event; MMRM = mixed model repeated measurements; Q4W = Every 4 weeks; SC = subcutaneous; TED = thyroid eye disease.

Note: Unless otherwise noted, all ocular efficacy analyses are for the study eye only.

^a For continuous endpoints, intercurrent events will follow a treatment policy strategy as the endpoints are not response based.

Table 2 Other Secondary Objectives and Endpoints

Other Secondary Objectives and Endpoints	
Efficacy Objectives	Corresponding Endpoints
To assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24 based on health-related quality of life	- Proportion of participants with active disease with a ≥ 6-point improvement in the Visual Functioning subscale of the GO-QoL from baseline at Week 24 - Proportion of participants with active disease with a ≥ 6-point improvement in the Appearance subscale of the GO-QoL from baseline at Week 24
To further assess the efficacy of subcutaneous satralizumab at Week 24	- Proportion of participants with active disease achieving overall response at Week 24, defined as a ≥ 2-point reduction in CAS, AND a ≥ 2 mm reduction in proptosis from baseline (Day 1) in the study eye, provided there is no corresponding deterioration in CAS or proptosis (≥ 2-point/mm increase) in the fellow eye - Proportion of participants with active disease achieving a ≥ 2-point reduction in CAS in the study eye from baseline (Day 1) to Week 24 - Proportion of participants with active disease achieving a CAS value of 0 or 1 in the study eye (indicating no or minimal inflammation, respectively) at Week 24 - Proportion of participants with a ≥ 10-point improvement in the OSDI overall scores across all levels of baseline severity at Week 24 in overall population - Change from baseline (Day 1) in the OSDI Ocular Symptoms, and Vision-related Function sub-scale scores at Week 24 in overall population - Change in Oxford corneal staining scores from baseline (Day 1) at Week 24^a in overall population - Proportion of participants achieving complete binocular diplopia response at Week 24 in overall population - Change from baseline (Day 1) in binocular diplopia at Week 24 in overall population

Table 2 Other Secondary Objectives and Endpoints (cont.)

To assess the efficacy of subcutaneous satralizumab at Week 48 in those participants who are proptosis non-responders at Week 24^b	- Proportion of participants achieving ≥ 2 mm reduction in proptosis at Week 48 in the study eye (provided there is no corresponding deterioration in proptosis [≥ 2 mm increase] in the fellow eye). - Proportion of participants achieving overall response at Week 48, defined as a ≥ 2-point reduction in CAS, AND a ≥ 2 mm reduction in proptosis from baseline (Day 1) in the study eye, provided there is no corresponding deterioration in CAS or proptosis (≥ 2-point/mm increase) in the fellow eye. - Proportion of participants achieving a ≥ 2-point reduction in CAS in the study eye from baseline (Day 1) at Week 48. - In the subset of participants with baseline (Day 1) diplopia (>0): Proportion of participants achieving Grade ≥ 1 reduction/improvement in diplopia at Week 48. - Proportion of participants with a ≥ 6-point improvement in the Visual Functioning subscale of the GO-QoL at Week 48. - Proportion of participants with a ≥ 6-point improvement in the Appearance subscale of the GO-QoL at Week 48. - Change in proptosis from baseline (Day 1) at Week 48 - Proportion of participants requiring surgical intervention for TED up to Week 48.
To assess the efficacy of subcutaneous satralizumab at Week 48 in those participants who are proptosis responders at Week 24^b	- Proportion of participants with worsening of proptosis by ≥ 2 mm from Week 24 to Week 48 - Proportion of participants with worsening of proptosis by ≥ 2 mm from baseline (day 1) to Week 48 - Proportion of participants with maintenance of proptosis response from Week 24 at Week 48. - Proportion of participants with maintenance of CAS response from Week 24 at Week 48 - Proportion of participants with maintenance of proptosis response from baseline (Day 1) at Week 48 - Proportion of participants with maintenance of CAS response from baseline (Day 1) at Week 48 - Change in proptosis from Week 24 to 48 - Change in CAS from Week 24 to 48 - Proportion of participants with a ≥ 6-point improvement in the Visual Functioning sub-scale score of the GO-QoL from Week 24 at Week 48 - Proportion of participants with a ≥ 6-point improvement in the Appearance sub-scale score of the GO-QoL from Week 24 at Week 48 - Change in proptosis from baseline (Day 1) to Week 48 - Change in CAS from baseline (Day 1) to Week 48 - Proportion of participants requiring surgical intervention for TED up to Week 48

Table 2 Other Secondary Objectives and Endpoints (cont.)

Safety Objective	Corresponding Endpoints
To evaluate the safety and tolerability of satralizumab	- Incidence, seriousness, and severity of AEs, with severity determined according to NCI CTCAE V5 - Change from baseline (Day 1) in vital signs, targeted clinical laboratory test results, ECG parameters, and body weight
Pharmacokinetic Objective ^c	Corresponding Endpoints
- To characterize the satralizumab PK profile by evaluating plasma exposure over 24 weeks - To evaluate potential relationships between drug exposure and the efficacy and safety of satralizumab	- Serum concentration of satralizumab at specified timepoints - Estimates of primary PK parameters (e.g., CL/F and V/F) and secondary PK parameters (e.g., AUC) derived using popPK modeling - Relationship between selected covariates and exposure to satralizumab - Relationship between serum concentration or PK parameters for satralizumab and efficacy endpoints, and safety endpoints

AE = adverse event; AUC = area under the concentration-time curve; CAS = clinical activity score; GO-QoL = Graves' Ophthalmopathy Quality of Life; CL = clearance; NCI CTCAE = National Cancer Institute Common Terminology Criteria for Adverse Events, Version 5.0; OSDI = ocular surface disease index; PK = pharmacokinetic; popPK = population pharmacokinetics; TED = thyroid eye disease; V/F = apparent volume of distribution.

^a The scale in the Oxford grading scheme chart categorizes corneal staining into 6 groups according to severity: 0 = absent; I = minimal; II = mild; III = moderate; IV = marked; and V = severe. The observer will compare the overall appearance of the participant's corneal staining with the reference figure in the protocol. The observer should select the appropriate grade that best represents the state of corneal staining. The staining score will be recorded for the exposed interpalpebral cornea and conjunctiva.

^b For the purposes of the Clinical Study Report, Part II efficacy results will be presented individually for Studies GP44467 and GP44729 and pooled across both studies.

^c The PK objectives will be described in another document.

Table 3 Exploratory Objectives and Endpoints

Exploratory Objectives and Endpoints	
Efficacy Objectives	Corresponding Endpoints
To further assess the efficacy of subcutaneous satralizumab compared with placebo at Week 24	- Change in CAS in participants with active disease from baseline (Day 1) to Week 24 - Proportion of participants in the overall population requiring orbital decompression surgery or any form of reconstructive surgery from baseline (Day 1) to Week 24 - Change in the motility component on the Clinical Measures of Severity scale (based on EUGOGO consensus) from baseline (Day 1) to Week 24 in the overall population - Proportion of participants with active disease with improvement for each component of the Clinical Measures of Severity scale (based on EUGOGO consensus) from baseline (Day 1) to Week 24 - Proportion of participants with active disease at each overall response level at Week 24 (high responders, responders, low responders and non-responders ^a) - Evaluation of each individual component of CAS at Week 24 in participants with active disease - In the subset of participants (both active and chronic inactive) with baseline (Day 1) orbital MRI, change in extraocular muscle and fat volume from baseline (Day 1) to Weeks 24 - Change in VA from baseline at Weeks 24 in the overall population
To assess the efficacy of subcutaneous satralizumab compared with placebo in participants with chronic inactive TED at Week 24	- Proportion of participants achieving ≥ 2 mm reduction in proptosis from baseline (Day 1) to Week 24 - Change in proptosis from baseline at Week 24 - Proportion of participants with a ≥ 6-point improvement in the Visual Functioning sub-scale score of the GO-QoL at Week 24 - Proportion of participants with a ≥ 6-point improvement in the Appearance sub-scale score of the GO-QoL at Week 24 - Proportion of participants achieving ≥ 1 grade reduction/improvement in diplopia at Week 24

Table 3 Exploratory Objectives and Endpoints (cont.)

To further assess the efficacy of subcutaneous satralizumab compared with placebo at Week 48	- In the subset of participants with active TED at baseline (Day 1): Time to first achieving CAS value of 0 or 1 in the study eye (indicating no or minimal inflammation, respectively) from baseline (Day 1) to Week 48 - Proportion of participants requiring orbital decompression surgery or any form of reconstructive surgery from baseline (Day 1) to Week 48 - Change in the motility component on the Clinical Measures of Severity scale (based on EUGOGO consensus) from baseline (Day 1) to Week 48 - Proportion of participants with improvement for each component of the Clinical Measures of Severity scale (based on EUGOGO consensus) from baseline (Day 1) to Week 48 - Proportion of participants at each overall response level at Week 48 (high responders, responders, low responders and non-responders ^a) - Evaluation of each individual component of CAS at Week 48 - In the subset of participants with baseline (Day 1) orbital MR imaging, change in extraocular muscle and fat volume from baseline (Day 1) to Week 48 - Change in VA from baseline at Week 48 - Proportion of participants with a ≥ 10-point improvement in the OSDI overall scores across all levels of baseline severity at Week 48 - Change from baseline (Day 1) in the OSDI Ocular Symptoms, and Vision-related Function subscale scores at Week 48 - Change in Oxford corneal staining scores from baseline (Day 1) at Week 48^b - Proportion of participants with proptosis response at Week 48 - Proportion of participants with a ≥ 6 point improvement in the Visual Functioning and Appearance subscale scores of the GO-QoL from baseline to Week 48
To further assess the efficacy of subcutaneous satralizumab compared with placebo at Week 72	Proportion of participants requiring medical therapy, orbital decompression surgery or any form of reconstructive surgery for TED from Week 48 through Week 72
Immunogenicity Objective	Corresponding Endpoint
To evaluate the immune response to satralizumab	Prevalence of ADAs at baseline (Day 1) and incidence of ADAs during the study

ADA=anti-drug antibody; AE=adverse event; AUC=area under the concentration-time curve; CAS=clinical activity score; EUGOGO=European Group on Graves Orbitopathy; GO-QoL=Graves' Ophthalmopathy Quality of Life; CL=clearance; IL-6=interleukin-6;

NCI CTCAE=National Cancer Institute Common Terminology Criteria for Adverse Events, Version 5.0; OSDI=ocular surface disease index; PK=pharmacokinetic; popPK=population pharmacokinetics; TED=thyroid eye disease; VA=visual acuity; V/F=apparent volume of distribution.

- ^a Participants will be classified into one of four response categories (CAS changes are measured on a 7-item scale) for the exploratory endpoint. Participants that meet criteria for several response categories will be allocated to the highest possible category.

High responders are participants who had a reduction in both proptosis (≥ 2 mm) and CAS (≥ 2 -points) from baseline (Day 1) in the study eye, and no deterioration in the fellow eye (i.e., increase in CAS ≥ 2 points or increase in proptosis ≥ 2 mm).

Responders are participants who had a reduction in either proptosis (≥ 2 mm) or (but not “and”) CAS (≥ 2 points) from baseline (Day 1) in the study eye (but no worsening in the other respective endpoint (proptosis or CAS as defined above in the study eye, and no deterioration in the fellow eye as defined above).

Low responders are participants who had a reduction in either proptosis (≥ 1 and < 2 mm) or (but not “and”) CAS (≥ 1 -point and < 2 -points) from baseline in the study eye (but no worsening in the other respective endpoint (proptosis or CAS as defined above in the study eye, and no deterioration in the fellow eye [as defined above]).

Non-Responders are participants who did not fit into any of the above categories

- ^b The scale in the Oxford grading scheme chart categorizes corneal staining into 6 groups according to severity: 0=absent; I=minimal; II=mild; III=moderate; IV=marked; and V=severe. The observer will compare the overall appearance of the participant’s corneal staining with the reference figure in the protocol. The observer should select the appropriate grade that best represents the state of corneal staining. The staining score will be recorded for the exposed interpalpebral cornea and conjunctiva.

1.5 STUDY DESIGN

This Phase III, randomized, double-masked, placebo-controlled, multicenter study is designed to evaluate the efficacy, safety, pharmacokinetics, and pharmacodynamics of satralizumab compared with placebo for the treatment of patients with moderate-to-severe active and chronic inactive TED.

The study includes the following phases:

- Screening period (up to 28 days long): During the screening period, individuals’ eligibility will be evaluated for study participation. The screening period may be extended in exceptional circumstances but cannot exceed 42 days.
- Placebo-controlled treatment periods: this will be split into two parts:
 - Part I: Weeks 1–24; with study treatment administered until Week 20, and primary endpoint assessed at Week 24.
 - Part II: Weeks 24–48, with study treatment administered until Week 44, and an analysis conducted at Week 48.

The final visit of Part I coincides with the first visit of Part II (Week 24). At this visit, response to treatment will be assessed. Proptosis non-responders from either arm will receive satralizumab treatment in Part II. Proptosis responders from either arm will be re-randomized to satralizumab or placebo in Part II, regardless of their original treatment assignment.

- Follow-up period: Weeks 48-72. Participants will be assessed every 2 months through telephone calls.

The primary endpoint will be assessed at Week 24, followed by an analysis at Week 48, and a final analysis will be conducted at Week 72. Protocol Figure 1 presents an overview of the study design. A schedule of activities is provided in the protocol, see Section 1.3.

1.5.1 Overview of Study Design

Approximately 120 participants who meet the study eligibility criteria will be randomized on Day 1 in a 1:1 ratio (stratified by current smoking status [current smoker, including the use of e-cigarettes vs. former smoker vs. never smoked] and TED disease activity status [active vs. inactive]) to receive seven SC doses of either satralizumab (with dosing based on body weight [REDACTED] or matching placebo in Part I. [REDACTED]
[REDACTED]

All participants will be allowed to receive rescue intervention per investigator's discretion from Week 4 through the end of the study per the following criteria:

- Diagnosis of sight-threatening disease for which prompt intervention is indicated in the opinion of the investigator, evidenced by:
 - Dysthyroid optic neuropathy, or
 - Corneal ulceration (defined as a deep corneal erosion which extends through epithelium and Bowman's membrane into corneal stroma)

Participants who receive rescue intervention between Weeks 4 and 48 will be discontinued from the study drug administration but will be encouraged to remain in the study to follow the protocol-defined schedule of activities, with a particular emphasis on the Week 24 and 48 visits, if possible. This will allow for monitoring of participants as well as collection of end-of-study data.

1.5.2 Part I Period

During Part I, eligible participants will be randomized to receive either satralizumab ([REDACTED] mg, [REDACTED] mg, or [REDACTED] mg based on body weight) or matching placebo on Day 1 and at Weeks 2, 4, 8, 12, 16, and 20, with a final visit of the Part I period at Week 24. The treatment which participants receive will be assigned via interactive voice or web-based response system (IxRS).

1.5.3 Part II Period

After the 24-week treatment period in Part I, participants will enter a second treatment period called Part II. The treatment which participants receive in Part II of the study will

be assigned via IxRS and is determined by their proptosis response (PR) at the end of Part I (Week 24 primary endpoint):

- Participants who are proptosis non-responders at Week 24 will receive satralizumab treatment during Part II (Weeks 24–44).
- Participants who are proptosis responders at Week 24 will be re-randomized in a 1:1 ratio (using the original randomization stratification factors measured at baseline [smoking status and TED disease activity status]) to receive SC satralizumab (with dosing based on body weight [REDACTED] or matching placebo (Weeks 24–44), regardless of their original treatment assignment in Part I.
 - Note:
 - The Week 24 visit must be recorded in IxRS for the re-randomization at Week 24 to occur. If the Week 24 visit is out of window, the subsequent visits in Part II will not align with the visit days listed in Protocol Table 1, since these are based on the Week 24 visit occurring on Day 169 $\pm$ 7 days.
 - In such cases where the participant's Week 24 visit is delayed, the site should use the IxRS visit window (and not the visit window in Protocol Table 1) for scheduling participant's subsequent visits in Part II, since the date of all subsequent visits in Part II would be based on the actual date of the Week 24 visit (Day 1 of Part II).
- Participants that discontinue the study drug in Part I but choose to stay in the trial for Part II will not be re-randomized and will not receive any further study treatment (satralizumab/placebo). These participants will continue to be followed in the study in order to monitor for safety and collect end of study data; during this time, they will be able to receive standard of care at the investigator's discretion.

During Part II of the study, the participants and the study staff will remain masked to the treatment the participants received during Part I of the study.

Due to the study design, participants who received placebo during Part I and receive satralizumab in Part II will receive a loading dose of satralizumab at Week 26; participants who received satralizumab in Part I and receive satralizumab in Part II do not require a loading dose and will receive placebo at Week 26.

The final visit at Week 48 in Part II (4 weeks after the final dose) is the last in-person visit of the study; subsequent follow up visits will be conducted by telephone (see Protocol Section 4.1.1.3).

1.5.4 Follow-up Period

After Week 48, participants will enter a 6-month telephone follow-up period. The participant will be contacted every 2 months (i.e., at Weeks 56, 64, and 72) for a telephone follow-up visit, to review adverse events, concomitant treatments, and any treatments or surgery received for TED.

If the participants choose to visit the study site in-person during the ‘telephone follow-up period’, the same information that would be collected from telephone conversation will be obtained; no other data will be recorded on the eCRF.

1.5.5 Treatment Assignment

Participants will be randomized in a 1:1 ratio (stratified by smoking status [current smoker, including the use of e-cigarettes vs. former smoker vs. never smoked] and TED disease activity status [active vs. inactive]) to receive seven SC doses of either satralizumab (with dosing based on body weight; see Protocol Table 4) or matching placebo on Day 1, at Weeks 2, and 4 (loading doses); then Q4W at Weeks 8, 12, 16, and 20 (maintenance doses), with a final visit at Week 24.

The treatment which participants receive in Part II of the study is then determined by their PR at the end of Part I (Week 24 primary endpoint):

- Participants who are proptosis non-responders at Week 24 will receive SC satralizumab during Part II (Weeks 24–44).
- Participants who are proptosis responders at Week 24 will be re-randomized in a 1:1 ratio (using the original randomization stratification factors measured at baseline (smoking status and TED disease activity status) to receive SC satralizumab or matching placebo (Weeks 24–44), regardless of their original treatment assignment in Part I.
- Participants that discontinue the study drug in Part I but choose to stay in the trial for Part II will not be re-randomized and will not receive any further study treatment (satralizumab/placebo [SAT/PLB]). These participants will continue to be followed in the study in order to monitor for safety and collect end of study data; during this time, they will be able to receive standard of care at the investigator’s discretion.

The final visit at Week 48 in Part II (4 weeks after the final dose) is the last onsite visit of the study. After Week 48, follow-up visits will be conducted every 8 weeks over telephone until Week 72 to review adverse events, concomitant treatments, and any treatments or surgery received for TED. During Part II of the study, the participants and the study staff will remain masked to the treatment the participants received during Part I of the study.

All participants will be allowed to receive rescue intervention per investigator’s discretion from Week 4 through the end of the study per rescue criteria (Protocol Section 6.1.2).

After randomization and at each study treatment visit, IxRS will assign the appropriate study treatment kit to be used.

1.5.6 Data Monitoring

An independent Data Monitoring Committee (iDMC) will be incorporated in this study.

[REDACTED]

The iDMC will monitor safety in an ongoing basis. Members of the iDMC will be external to the Sponsor and will follow a charter that outlines the committee's roles and responsibilities. [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

Full details regarding the roles and responsibilities of the iDMC will be provided in the charter.

[REDACTED] for each time the iDMC conducts an unmasked interim safety review prior to the primary analysis. At the time of the primary analysis, it is estimated that [REDACTED] data reviews will have been conducted by the iDMC; therefore, efficacy analyses will be performed at a significance level of [REDACTED]. The actual adjustment will depend on the actual number of iDMC interim safety reviews.

2. **ANALYSIS SETS**

The participant analysis sets for the purposes of analyses are defined in [Table 4](#).

Table 4 Participant Analysis Sets

Participant Analysis Set	Description
Part I FAS	Include all participants randomized; participants will be included in the analyses according to the treatment they were assigned at baseline (Study Day 1).
Part I SAS	Include all participants exposed to study treatment; participants will be analyzed according to the treatment that they actually received.
Part II FAS	Include participants who entered Part II and were re-randomized on the Week 24 visit; participants will be included in the analyses according to the treatment they were assigned to at the Week 24 visit (re-randomization). Non-responders at Week 24 and participants that enter Part II off-drug are not included.
Part II non-randomized	Include participants who were non-responders in Part I and entered Part II on treatment.
Part II all cases	Include participants that entered Part II. Participants will be included in the analyses according to the treatment to which they were assigned at the Week 24 visit, irrespective of response. Participants that entered Part II off treatment will be counted as placebo.

Participant Analysis Set	Description
Part II off treatment	Include participants that entered Part II off treatment.
Part II SAS	Include all participants who entered Part II and were exposed to study treatment (in any part of the trial-); participants will be analyzed according to the different possible treatment combinations they actually received in Part I and II. No distinction between responders and non-responders is made.
Satralizumab SAS	All participants exposed at least once to Satralizumab (in any part of the trial).

FAS=full analysis set; SAS=safety analysis set

The analysis sets will be used for both analyses on active participants only, and for analysis on the overall population (active and chronic participants combined).

3. STATISTICAL HYPOTHESES AND SAMPLE SIZE DETERMINATION

3.1 STATISTICAL HYPOTHESES

The primary efficacy endpoint is PR, defined as the proportion of participants achieving a ≥ 2 mm reduction in proptosis from baseline (Day 1) to Week 24 in the study eye, provided there is no corresponding deterioration of proptosis (≥ 2 mm increase) in fellow eye in the active full analysis set (FAS). Participants receiving rescue therapy, prohibited therapy, orbital or ocular surgery/procedure for TED are evaluated as non-responders, irrespective of early withdrawal of study treatment. The following hypothesis will be tested:

- Superiority of satralizumab every four weeks (Q4W) versus placebo in PR in the Part I FAS (active population).

The null and alternative hypothesis for the superiority test are as follows:

- The null hypothesis (H_0) is: $P_{\text{satralizumab}} - P_{\text{placebo}} = 0$
- The alternative hypothesis (H_a) is: $P_{\text{satralizumab}} - P_{\text{placebo}} \neq 0$

where $P_{\text{satralizumab}}$ and P_{placebo} are the expected PR for the Satralizumab arm and placebo arm, respectively.

If the observed rate is higher in the treatment arm and the two-sided p-value < 0.05 (), the test is considered positive and satralizumab is considered superior to placebo.

A fixed-sequencing test approach will be used to control the overall type I error rate for the hypothesis testing of the primary and key secondary endpoints.

If the primary endpoint test is positive, superiority of satralizumab compared with placebo for the following key secondary endpoints will be tested at the two-sided significance level mentioned above in the order as shown below:

- Proportion of participants achieving PR from baseline (Day 1) at Week 24 in the Part I FAS (overall population)
- Change in proptosis from baseline (Day 1) at Week 24 in the Part I FAS (active population)
- Proportion of participants with Grade ≥ 1 reduction/improvement in diplopia from baseline (Day 1) at Week 24 in the Part I FAS (active population) among participants with diplopia present at baseline
- Change in proptosis from baseline (Day 1) to Week 24 in the Part I FAS (overall population)
- Proportion of participants with Grade ≥ 1 reduction/improvement in diplopia from baseline (Day 1) at Week 24 in the Part I FAS (overall population) among participants with diplopia present at baseline
- Proportion of participants achieving absence of motility-induced pain from baseline (Day 1) at Week 24 in the Part I FAS (active population) among participants with motility-induced pain present at baseline
- Proportion of participants achieving absence of spontaneous pain from baseline (Day 1) at Week 24 in the Part I FAS (active population) among participants with spontaneous pain present at baseline

For hypotheses that use binary endpoints, the corresponding estimand framework will follow the primary estimand framework.

The continuous key secondary endpoints will be tested using the mixed-effect model for repeated measures method. The respective estimand framework can be found in [Table 1](#). If the primary endpoint is met in both trials, the hypotheses for Part II will be tested on the pooled study population (from studies GP44467 and GP44729) according to the graphical approach by ([Bretz and Xi 2019](#)), see Section 4.6.

3.1.1 Analyses for the European Medicines Agency

To support registration in the European Union, a separate set of hypotheses will be tested in the active and overall population after the primary hypothesis. For the European Medicines Agency, the following hypotheses will be tested hierarchically at an overall significance level of $\alpha=0.05$ (), to control the overall type I error rate:

-

- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]

Intercurrent events will be defined as in the primary estimand definition. For binary endpoints, intercurrent events will be handled as for the primary estimand. For continuous endpoints, a treatment policy approach will be used for handling all intercurrent events as the endpoints are not response based.

3.1.2 Analyses for the Japan Pharmaceuticals and Medical Devices Agency

To support registration in Japan, a separate set of hypotheses will be tested in the active and overall population after the primary hypothesis. For the Pharmaceuticals and Medical Devices Agency, the following hypotheses will be tested hierarchically at an overall level of $\alpha=0.05$ ([REDACTED]), to control the overall type I error rate:

- [REDACTED]

- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]

3.2 SAMPLE SIZE DETERMINATION

A total of 120 participants including 25% chronic inactive participants will be enrolled in this study and randomized at baseline in a 1:1 ratio (stratified by current smoking status [current smoker, including the use of e-cigarettes vs. former smoker vs. never smoked] and TED disease activity status [active vs. inactive]).

A total of 90 participants with active TED will provide > 90% power to show superiority of satralizumab compared with placebo in PR at Week 24. The following assumptions are used:

- Response ratio of [REDACTED] in the satralizumab arm
- Response ratio of [REDACTED] in the placebo arm
- Fisher exact test for two proportions
- 5% two-sided type I error rate
- 10% dropout rate

While the Mantel-Haenszel (MH) test will be used as the main analysis method, the Fisher exact test is expected to be less sensitive and a more conservative way to estimate the sample size.

A total of 120 active and chronic inactive participants (Part I FAS [overall population]) will provide > 90% power to show superiority of satralizumab compared with placebo in PR at Week 24. The following assumptions are used:

- Response ratio of ■■■ in the satralizumab arm for active participants
- Response ratio of ■■■ in the placebo arm for active participants
- An unknown response rate in the satralizumab and placebo arm for inactive participants, but with a difference in response rates of ■■■.
- Fisher exact Test for two proportions
- 5% two-sided type I error rate
- 10% dropout rate

4. STATISTICAL ANALYSES

The primary analysis will be performed when all participants have either completed the study through Week 24 or have discontinued from the study prior to Week 24, whichever comes later (Part I). An update analysis will be performed when all participants have either completed the study through Week 48 or have discontinued from the study prior to Week 48; whichever comes later (Part II).

The final analysis database lock will be performed when all participants have either completed the study through Week 72 (last participant last visit) or have discontinued early from the study, whichever comes later. The analyses specified in this statistical analysis plan (SAP) supersede those specified in the protocol.

4.1 GENERAL CONSIDERATIONS

Unless otherwise noted, efficacy analyses and participant disposition will be based on the FAS population and grouped according to the treatment to which they were assigned. Safety analyses will be performed based on safety analysis set (SAS) populations and grouped according to the treatment that they received. Immunogenicity analyses will be done on the immunogenicity analysis set.

Binary efficacy endpoints will be analyzed by evaluating the difference of proportions.

Continuous efficacy endpoints will be analyzed by calculating the least square means using mixed model repeated measurements (MMRM).

For the primary analysis, only data from Part I will be included. Efficacy and safety analyses will be provided for GP44467 and GP44729 respectively.

For the Week 48 analysis, data from both Part I and Part II will be included. Efficacy and safety analyses will be provided for GP44467 and GP44729 respectively. For the purposes of the Clinical Study Report, Part II efficacy results will be presented individually for Studies GP44467 and GP44729 and pooled across both studies.

For the Week 72 analysis, only data from the follow-up period will be included. All analyses will be provided for GP44467 and GP44729 respectively.

The baseline for efficacy analyses in Part I is defined as the last available measurement prior to first exposure to study treatment on Study Day 1. Typically, the baseline will be the pre-dose Day 1 assessments but may include screening results if the pre-dose Day 1 result is not available or missing, or if it is not collected at Day 1 according to the schedule of assessments. In case there was no exposure to study drug, the last available measurement prior to randomization in Part I will be used.

Baseline for safety analyses in Part I is defined as the last available measurement prior to first exposure to study drug. Participants with missing baseline assessments will not be imputed.

Baseline for efficacy analyses in Part II is defined as the last available measurement prior to receiving the study drug on the Week 24 visit. In case there was no exposure to study drug in Part II, the last available measurement prior to re-randomization in Part II will be used. If the last available measurement happened before Week 16 then the baseline value for Part II will be set to be missing.

Baseline for safety analyses in Part II is defined as the last available measurement prior to administration of the study drug in Part II/Week 24 assessment. Participants with missing baseline assessments will not be imputed.

Unless otherwise noted, analyses of efficacy outcome measures relating to Part I will be stratified by the randomization stratification factors assessed at baseline Part I (Study Day 1): current smoking status (current smoker, including the use of e-cigarettes vs. former smoker vs. never smoked) and TED disease activity status (active vs. inactive). Efficacy analyses for Part II will be stratified by the same randomization stratification factors assessed at baseline Part I (Study Day I).

Unless otherwise stated, efficacy analyses relating to Part II will use data from Week 24 (re-randomization) to Week 48 if they are comparative in nature.

Analyses that use data from Day 1 to Week 48 or beyond will group participants by treatment combinations and be non-comparative.

For the primary analysis, [REDACTED] [REDACTED] for each time the iDMC conducts an unmasked interim safety review prior to the primary analysis. At the time of the primary analysis, it is estimated that [REDACTED] data reviews will have been conducted by the iDMC; therefore, it is planned that the efficacy analyses will be performed at a significance level of [REDACTED]. The actual adjustment will depend on the actual number of iDMC interim safety reviews.

The estimates and CIs will be provided for the difference in means or proportions between two treatment groups, and for the hazard ratio (for time-to-event variables). All confidence intervals (CI) will be two-sided and at the [REDACTED] level (assuming [REDACTED] data reviews).

Descriptive summaries will include the mean, standard deviation (SD), median, and range for continuous variables, and counts and percentages for categorical variables. Missing data for descriptive summaries will not be imputed.

Data collected in scheduled and unscheduled visits will be mapped to visits that appear in the schedule of assessments per the protocol using the actual study day of assessment. However, no visit happening in Part I will be mapped to Part II and vice versa. If there are multiple values in the same window, values closest to the targeted study day for the scheduled assessment will be used in the analysis if not specified otherwise (e.g. safety values). In the case of a tie, the value with the latest datetime will be selected.

The value to be used in the analysis will be selected prior to the implementation of intercurrent event strategies.

For the FAS populations, the day of first dose after randomization will be designated Study Day 1. Each assessment will be assigned a study day calculated as: *date of assessment - date of first dose after randomization + 1*.

4.2 PRIMARY ESTIMAND ANALYSIS

4.2.1 Definition of Primary Endpoint/Estimand

The primary efficacy endpoint is the proportion of participants achieving ≥ 2 mm reduction in proptosis from baseline (Day 1) at Week 24 in the study eye, provided there is no deterioration of proptosis (≥ 2 mm increase) in the fellow eye, as defined in [Table 1](#).

The primary estimand is defined as follows:

- Population: adult participants with moderate to severe active thyroid eye disease. The population corresponds to the Part I FAS (active population).
- Variable: proportion of participants achieving ≥ 2 mm reduction in proptosis from baseline (Day 1) at Week 24 in the study eye, provided there is no deterioration of proptosis (≥ 2 mm increase) in the fellow eye.
- Treatment (as randomized):
 - Experimental: satralizumab SC [REDACTED] mg, [REDACTED] mg, or [REDACTED] mg (for bodyweight [REDACTED] respectively), at Day 1, Week 2, and Week 4 (loading doses) and maintenance doses Q4W thereafter until Week 20. Seven doses in total during the 24-week Part I period.

- Control: placebo SC at Day 1, Week 2, and Week 4 (loading doses), and maintenance doses Q4W thereafter until Week 20. Seven doses in total during the 24-week Part I period.
- Intercurrent events and handling:

ICE	ICE Strategy	Data used
Early discontinuation of study treatment due to AEs or lack of efficacy	Treatment Policy	All observed values will be used regardless of the occurrence of the ICE
Use of any prohibited systemic treatment or prohibited therapy for either eye (study eye or fellow eye):	Composite variable strategy	Participants with this intercurrent event will be regarded as non-responders
Administration of rescue treatment	Composite variable strategy	Participants with this intercurrent event will be regarded as non-responders
Orbital or ocular surgery/procedure for TED (e.g., orbital decompression, eyelid or strabismus surgery) in either eye (study eye or fellow eye)	Composite variable strategy	Participants with this intercurrent event will be regarded as non-responders

AE = adverse event; ICE = intercurrent event; TED = thyroid eye disease.

In the case of several intercurrent events for the same participant, the composite variable strategy will take priority, i.e., the participant will be regarded as a non-responder from the earliest time of an ICE with a composite strategy.

- Population-level summary: difference in proportions of proptosis responders at Week 24 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing MH stratum weights.

4.2.2 **Main Analytical Approach for the Primary Endpoint Analysis Method (Assuming No Missing Data)**

The population difference in the proportions of proptosis responders at Week 24 will be estimated as a weighted mean of stratum specific risk differences utilizing MH stratum weights ([Mantel and Haenszel, 1959](#)).

Specifically, the MH estimate of the common risk difference is defined as

$$\hat{d}_{MH} = \sum_{\text{Strata } h} \hat{d}_h \cdot w_h.$$

$\hat{d}_h$ is the observed risk difference in stratum h (i.e. the difference in the observed proportion of responders in the intervention and control arms in that stratum), and w_h is the stratum weight:

$$w_h = \frac{(n_{h1}n_{h2})/n_h}{\sum_{\text{Strata } i} (n_{i1}n_{i2})/n_i}$$

where n_{i1} and n_{i2} denote the intervention and control arm sample size of stratum i , and n_i is the total sample size of the stratum. The standard error $\hat{\sigma}(\hat{d}_{MH})$ of $\hat{d}_{MH}$ can be estimated using the method proposed by Sato (1989).

The hypothesis test of the null hypothesis that the true difference in the probability of a response between the two treatment groups is 0 versus the alternative that it is different from 0 will be based on the Z-statistic:

$$Z = \hat{d}_{MH} / \hat{\sigma}(\hat{d}_{MH})$$

which has an approximate standard normal distribution under the null hypothesis.

A corresponding two-sided 95% CI for the difference in the probability of a response between the two treatment groups is defined as:

$$\hat{d}_{MH} \pm z_{0.975} \cdot \hat{\sigma}(\hat{d}_{MH})$$

where $z_{0.975}$ is the 97.5% quantile of the standard normal distribution.

Handling of Missing Data

Missing proptosis values in the study and the fellow eye, respectively, at all follow-up visits until Week 24 will be multiply imputed prior to the derivation of the primary endpoint. Note that continuous proptosis measurements, not the PR (which is a derived variable) will be imputed. Note also that only imputed proptosis values at Week 24 in subjects without intercurrent events (ICE) handled using a composite strategy will actually affect the derived primary endpoint.

Missing proptosis values in the study and fellow-eye at all follow-up visits until Week 24 will be multiply imputed [REDACTED]. Specifically, the imputation uses sequential (monotone) imputation (van Buuren 2008) as described below:

1. [REDACTED]
2. [REDACTED]

Analysis and Inference Following Imputation of Missing Data

For each imputed dataset, the primary composite endpoint will be derived. Treatment effect estimates (i.e., MH estimates of the common risk difference) and corresponding standard errors will be combined across the imputed datasets using Rubin's rules ([Rubin 1987](#)). Based on these, the primary hypothesis test and a corresponding 95% CI of the common risk difference will be obtained as described for the complete data case.

4.2.3 Sensitivity Analyses

A sensitivity analysis will be performed using the same intercurrent events, handling of intercurrent events and analysis methods as the main analysis of the primary endpoint, with the exception that missing data will not be imputed and participants with missing data will be excluded from the analysis.

Furthermore, a tipping point analysis will be performed to explore the robustness of the primary endpoint:

- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]
- [REDACTED]

Furthermore, the treatment effect will be assessed as an odds ratio (satralizumab response odds vs placebo response odds) using a logistic regression model. The model will be adjusted for the randomization stratification factor (smoking status).

[REDACTED] A sensitivity analysis for the primary estimand will be conducted treating all patients from this site as non-responders.

4.2.4 Supplementary Analyses

A supplementary analysis will be performed using the same intercurrent events, analysis methods and missing data handling approach as the main analysis of the primary

endpoint, with the exception that participants with any intercurrent event will be treated as non-responders.

Furthermore, the percentage change from baseline to Week 24 in proptosis in the Part I FAS will be analyzed and compared across arms in Part I.

Also, the analysis of the primary estimand will be repeated on the Part I FAS using a different criteria for PR, [REDACTED].

[REDACTED]
[REDACTED] A supplementary analysis for the primary endpoint will be conducted excluding patients from this site.

Subgroup Analyses for Primary Endpoint

In order to assess the robustness and generalizability of the primary estimand results when comparing satralizumab to placebo, the treatment effect is estimated in the following subgroups (using the assessment at baseline Part I):

- Smoking Status (current smoker, including the use of e-cigarettes, former smoker, never smoked)
- Age group (< 65 years, ≥ 65 years)
- Gender (female, male)
- [REDACTED] mg vs. [REDACTED] mg satralizumab (based on patients' bodyweight)
- Race (Asian, Black or African American, Native Hawaiian or other Pacific Islander, White, Not reported/unknown)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino, Not reported/unknown)
- Iris pigmentation (dark/light)

Summaries of the primary estimand by these subgroups will be provided in forest plots.

The analysis will be performed within each subgroup separately. The estimand and data handling rules follow those for the main analysis (see Section 4.2.2). Point estimates of treatment group differences in PR with 95% CIs will be provided in forest plots. The subgroup categories may be combined if there is not enough representation of a specific subpopulation.

4.3 STRATIFICATION FACTORS

The analyses for the primary, secondary and exploratory will be stratified by smoking status, categorized as:

- Current smoker (including the use of e-cigarettes)
- Former smoker
- Never smoked

Analyses on the overall population will additionally be stratified by disease status (active, chronic inactive).

The stratification variables will be taken from the IxRS.

4.4 SECONDARY ENDPOINTS

4.4.1 Key Secondary Endpoints

The set of key secondary endpoints can be found in [Table 1](#).

For hypotheses that use binary endpoints, the corresponding estimand framework will follow the primary estimand framework, i.e. the same estimator, intercurrent events and handling strategies for intercurrent events will be used.

The continuous key secondary endpoints will be tested using the mixed-effect model for repeated measures method. The respective estimand framework can be found in [Table 2](#). The dependent variable corresponds to the change from baseline of the respective outcome variable. Unless otherwise specified, the MMRM model will include the continuous outcome variable of interest as the response variable, and will include the following fixed-effect terms:

- Treatment group as categorical variable
- Study visit as categorical variable. If needed, (e.g., if programming is done in R) the study visit should be defined as a factor with levels ordered according to the study week
- Visit-by-treatment group interaction
- Continuous values of the outcome variable at baseline
- Randomization stratification factors (disease status and smoking status) at baseline (Study Day 1)

An unstructured covariance matrix will be the first choice to fit the MMRM.

If MMRM with an unstructured covariance matrix does not converge, it will be replaced with a first order ante-dependence pattern and then with Toeplitz with heterogeneous variances, in that order.

An equal weight least square means (ls_m) assessing the adjusted difference of treatment group means will be calculated and used for the corresponding hypothesis test:

The null and alternative hypothesis for the corresponding superiority test are as follows:

- The null hypothesis (H_0) is: $ls_{m \text{ satralizumab} - \text{placebo}} = 0$
- The alternative hypothesis (H_a) is: $ls_{m \text{ satralizumab} - \text{placebo}} \neq 0$

4.4.1.1 Handling of Missing Data for Key Secondary Endpoints:

Missing data for key secondary endpoints will not be imputed separately, but implicitly by using MMRM. For binary endpoints, missing data will be evaluated as non-response.

No multiple imputation methods will be applied.

4.4.2 Secondary Endpoints

The set of secondary endpoints can also be found in [Table 2](#).

Unless otherwise specified, for analyses that use binary endpoints, the corresponding estimand framework will follow the primary estimand framework.

The continuous secondary endpoints will be tested using the mixed-effect model for repeated measures method. The respective estimand framework can be found in [Table 2](#). The methodology of the MMRM will align with the key secondary endpoints.

4.4.2.1 Handling of Missing Data for Secondary Endpoints:

Missing data for key secondary endpoints will not be imputed. For binary endpoints, missing data at Week 24 will be evaluated as non-response.

For continuous endpoints, missing values will be handled through the MMRM method.

4.4.2.2 Binary endpoints

The proportion of participants with a ≥ 6 -point improvement in the Visual Functioning and Appearance sub-scale of the GO-QoL from baseline at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS, restricted to participants with an overall sub-scale of ≤ 94 at Day 1 respectively.

The proportion of participants with active disease achieving overall response at Week 24, (defined as a ≥ 2 -point reduction in CAS, AND a ≥ 2 mm reduction in proptosis from baseline (Day 1) in the study eye, provided there is no corresponding deterioration in CAS or proptosis (≥ 2 -point/mm increase) in the fellow eye) will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS (active TED participants only).

The proportion of participants with active disease achieving a ≥ 2 -point reduction in CAS in the study eye from baseline (Day 1) to Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS (active TED participants only).

The proportion of participants achieving a CAS value of 0 or 1 in the study eye (indicating no or minimal inflammation, respectively) at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS (active TED participants only).

The proportion of participants with a ≥ 10 -point improvement in the ocular surface disease index (OSDI) overall scores across all levels of baseline severity at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS, restricted to participants with a score of ≥ 10 at Day 1.

The proportion of participants achieving complete binocular diplopia response (diplopia score=0) at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS, restricted to participants with active disease and with baseline diplopia > 0 .

The Change from Baseline at Week 24 in binocular diplopia will be categorized into the following five categories:

- Significant improvement: change from Baseline of -2 or -3.
- Improvement: change from Baseline of -1.
- No change: 0.
- Worsening: change from Baseline of 1.
- Significant worsening: change from Baseline of 2 or 3.

A proportional odds logistic regression model will be used to analyze the change from Baseline at Week 24 in binocular diplopia at Week 24 based on the ordinal response categories described above. The results will be compared across arms using an averaged odds ratio. The analysis will be done on the active Part I FAS.

The proportion of participants with worsening of proptosis by ≥ 2 mm from Week 24 to Week 48 will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS.

The proportion of participants with worsening of proptosis by < 2 mm in proptosis at Week 48 from Week 24 in the study eye (provided there is no corresponding deterioration in proptosis [≥ 2 mm increase] in the fellow eye) will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS.

The proportion of participants with a ≥ 6 -point improvement in the Visual Functioning and Appearance sub-scales of the GO-QoL from Week 24 at Week 48 will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS, restricted to participants with a GO-QoL sub-scale of ≤ 94 at Week 24 respectively.

4.4.2.3 Continuous Endpoints:

The change from baseline (Day 1) in the OSDI Ocular Symptoms, and Vision-related Function sub-scale scores at Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

The change in Oxford corneal staining scores from baseline (Day 1) at Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

The change in proptosis from Week 24 to 48 will be calculated and compared across (randomized) arms using MMRM. The analysis will be done on the Part II FAS.

The change in CAS from Week 24 to 48 will be calculated and compared across (randomized) arms using MMRM. The analysis will be done on the Part II FAS (active TED participants).

4.4.2.4 Non-Comparative Analyses

The proportion of participants achieving overall response at Week 48, defined as a ≥ 2 -point reduction in CAS, AND a ≥ 2 mm reduction in proptosis from baseline (Day 1) in the study eye, provided there is no corresponding deterioration in CAS or proptosis (≥ 2 -point/mm increase) in the fellow eye) will be calculated for every treatment combination occurring in Part I and II. The proportions will NOT be compared across arms. The analysis will be done separately on the Part II non-randomized (receiving SAT/SAT and PLB/SAT).

In the subset of participants with baseline (Day 1) diplopia (>0): The proportion of participants achieving Grade ≥ 1 reduction/improvement in diplopia from baseline to Week 48 will be calculated for every treatment combination occurring in Part I and II. The proportions will NOT be compared across arms. The analysis will be done on the Part II non-randomized (receiving SAT/SAT and PLB/SAT).

The proportion of participants with a ≥ 6 -point improvement from Day 1 in the Visual Functioning and Appearance sub-scales of the GO-QoL at Week 48 will be calculated for every treatment combination occurring in Part I and II. The proportions will NOT be compared across arms. The analysis will be done on the Part II all cases, restricted to participants with a ≤ 94 sub-scale at Study Day 1 respectively.

The proportion of participants achieving a ≥ 2 -point reduction in CAS in the study eye from baseline (Day 1) to Week 48 will be calculated for every treatment combination occurring in Part I and II. The proportions will NOT be compared across arms. The analysis will be done on the active Part II FAS.

The mean change in proptosis from baseline (Day 1) to Week 48 will be calculated for every treatment combination occurring in Part I and II using MMRM. The changes will NOT be compared across arms. The analysis will be done on the Part II all cases.

The proportion of participants with worsening of proptosis by ≥ 2 mm from baseline (Day 1) to Week 48 will be calculated for every treatment combination occurring in Part I and II. The proportions will NOT be compared across arms. The analysis will be done on the Part II all cases.

The proportion of participants requiring surgical intervention for TED from Day 1 to Week 48 will be calculated for every treatment combination occurring in Part I and II. The proportions will NOT be compared across arms. The analysis will be done on the Part II all cases.

The mean change in proptosis from baseline (Day 1) to Week 48 will be calculated for every treatment combination occurring in Part I and II using MMRM. The changes will NOT be compared across arms. The analysis will be done on the Part II non-randomized (receiving SAT/SAT and PLB/SAT)

Change in CAS from baseline (Day 1) to Week 48 will be calculated for every treatment combination occurring in Part I and II using MMRM. The proportions will NOT be compared across arms. The analysis will be done on the Part II all cases (active TED participants).

4.5 EXPLORATORY ENDPOINT ANALYSIS

4.5.1.1 Handling of Missing Data for Exploratory Endpoints:

Missing data for exploratory endpoints will not be imputed. For binary endpoints, missing data will be evaluated as non-response.

For continuous endpoints, missing values will be handled through the MMRM method.

4.5.1.2 Binary Endpoints

The proportion of participants requiring orbital decompression surgery or any form of reconstructive surgery for TED in the study eye from baseline (Day 1) to Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS.

The proportion of participants with improvement for each component of the Clinical Measures of Severity scale (based on European Group on Graves' Orbitopathy [EUGOGO] consensus) from baseline (Day 1) to Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS. Participants that cannot improve because of e.g., the best possible score at baseline will be considered to have improved if they maintain the baseline assessment at the Week 24 visit for all categorical items except subjective diplopia.

The proportion of participants at each overall response level at Week 24 (high responders, responders, low responders and non-responders, see definition below) will

be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS.

Participants will be classified into one of four response categories (CAS changes are measured on a 7-item scale). Participants that meet criteria for several response categories will be allocated to the highest possible category.

High responders are participants who had a reduction in both proptosis (≥ 2 mm) and CAS (≥ 2 -points) from baseline (Day 1) in the study eye, and no deterioration in the fellow eye (i.e., increase in CAS ≥ 2 -points or increase in proptosis ≥ 2 mm).

- Responders are participants who had a reduction in either proptosis (≥ 2 mm) or (but not “and”) CAS (≥ 2 -points) from baseline (Day 1) in the study eye (but no worsening in the other respective endpoint (proptosis or CAS as defined above in the study eye, and no deterioration in the fellow eye [as defined above])).
- Low responders are participants who had a reduction in either proptosis (≥ 1 and < 2 mm) or (but not “and”) CAS (≥ 1 -point and < 2 -points) from baseline in the study eye (but no worsening in the other respective endpoint (proptosis or CAS as defined above in the study eye, and no deterioration in the fellow eye [as defined above])).
- Non-responders are participants who did not fit into any of the above categories.

Each individual component of CAS at Week 24 will be evaluated. For each component, the proportion of participants with a response corresponding to “0” will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS, restricted to participants that score “1” at study Day 1 respectively.

The proportion of participants achieving ≥ 2 mm reduction in proptosis from baseline (Day 1) at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS (inactive TED participants only).

The proportion of participants with a ≥ 6 -point improvement in the overall Visual Functioning and Appearance sub-scales of the GO-QoL at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS (inactive TED participants only), restricted to participants with a subscale ≤ 94 at Study Day 1 respectively.

The proportion of participants achieving ≥ 1 grade reduction/improvement in diplopia at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS (inactive TED participants only) with a diplopia score > 0 at Study Day 1.

In participants with bilateral disease, the proportion of participants with a PR in the fellow eye at Week 24 will be calculated for each treatment group in Part I and compared across arms. The analysis will be done on the Part I FAS.

4.5.1.3 Continuous Endpoints

The change from baseline (Day 1) to Week 24 in CAS will be calculated and compared across arms using MMRM. The analysis will be done on the active Part I FAS.

The change in the motility component on the Clinical Measures of Severity scale (based on EUGOGO consensus) from baseline (Day 1) to Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

In the subset of participants with baseline (Day 1) orbital MRI, the change in proptosis from baseline (Day 1) to Week 24 as measured per MRI assessment will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

In the subset of participants with baseline (Day 1) orbital MRI, the change in extraocular muscle and fat volume from baseline (Day 1) to Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

The change in visual acuity (VA) (transformed to log MAR scale) from baseline to Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

Invalid VA (VA values that are confirmed to be inaccurate, due to the VA test being performed incorrectly) will be excluded from the analyses and provided in a separate listing. Non-standard VA data (such as VA test performed by uncertified VA assessor) will be included in the analyses.

The change in intraocular pressure (IOP) from baseline to Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS.

The change in proptosis from baseline to Week 24 will be calculated and compared across arms using MMRM. The analysis will be done on the Part I FAS (inactive TED participants).

4.5.1.4 Non-Comparative Analyses

All non-comparative analyses will evaluate data from baseline (Day 1) to Week 48 for every treatment combination occurring in Part I and II. The results will not be compared across arms and no p-values will be provided.

The time to first achieving CAS value of 0 or 1 in the study eye (indicating no or minimal inflammation, respectively) will be calculated using Kaplan Meier curves. The analysis will be done on the Part II all cases – active TED participants only (with respect to Day 1).

The proportion of participants requiring orbital decompression surgery for TED in the study eye or any form of reconstructive surgery will be calculated. The analysis will be done on the Part II all cases.

The change in the motility component on the Clinical Measures of Severity scale (based on EUGOGO consensus) will be calculated using MMRM. The analysis will be done on the Part II all cases.

For each component of the Clinical Measures of Severity scale (based on EUGOGO consensus), the proportion of participants with respective will be calculated. The analysis will be done on the Part II all cases, restricted to participants that can improve based on their baseline assessment.

The proportion of participants at each overall response level at Week 48 (high responders, responders, low responders and non-responders (see definition above) will be calculated. The analysis will be done on the Part II all cases.

The proportion of participants with proptosis response at Week 48 will be calculated. The analysis will be done on the Part II all cases. The proportion of participants with a ≥ 6 -point improvement in the Visual Functioning and Appearance subscales of the GO-QoL from baseline to Week 48 will also be calculated.

Each individual component of CAS at Week 48 will be evaluated. For each component, the proportion of participants with a response corresponding to “0” will be calculated. The analysis will be done on the Part II cases, restricted to participants scoring “1”, at Study Day 1.

The change in extraocular muscle and fat volume using MMRM. The analysis will be done on the Part II all cases-subset of participants with baseline (Day 1) orbital MR imaging.

The change VA (transformed to logMAR scale) will be calculated using MMRM. The analysis will be done on the Part II all cases.

Invalid VA (VA values that are confirmed to be inaccurate, due to the VA test being performed incorrectly) will be excluded from the analyses and provided in a separate listing. Non-standard VA data (such as VA test performed by uncertified VA assessor) will be included in the analyses.

The proportion of participants with a ≥ 10 -point improvement from Day 1 in the OSDI overall scores across all levels of baseline severity at Week 48 will be calculated. The analysis will be done on the Part II all cases, restricted to participants with a score of ≥ 10 at Day 1.

The change VA (transformed to logMAR scale) will be calculated using MMRM. The analysis will be done on the Part II all cases.

The change from baseline (Day 1) to Week 48 in the OSDI Ocular Symptoms and

Vision-related Function sub-scale scores will be calculated using MMRM. The analysis will be done on the Part II all cases.

The change in Oxford corneal staining scores will be calculated using MMRM. The analysis will be done on the Part II all cases.

The scale in the Oxford grading scheme chart categorizes corneal staining into 6 groups according to severity: 0=absent; I=minimal; II=mild; III=moderate; IV=marked; and V=severe. The observer will compare the overall appearance of the participant's corneal staining with the reference figure in Protocol Appendix 10. The observer should select the appropriate grade that best represents the state of corneal staining. The staining score will be recorded for the exposed interpalpebral cornea and conjunctiva.

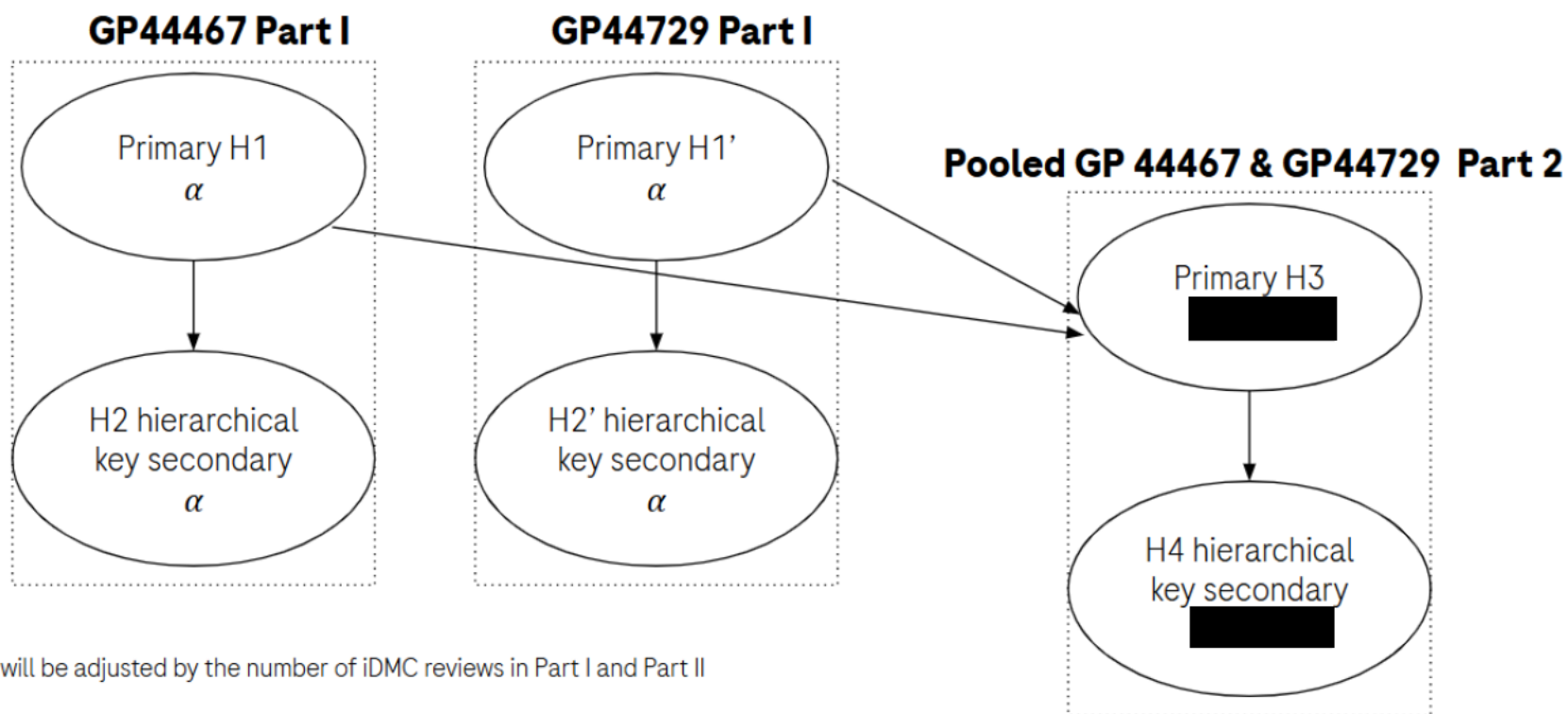
4.6 HYPOTHESES TESTING FOR PART II

At the Week 48 analysis, a separate set of hypotheses will be tested as long as the primary endpoint is met in both trials. The type I error-controlled testing will be performed on the pooled set of both studies (GP44467 and GP44729), with the study being adjusted for in the analysis. Missing data will be evaluated as non-response/loss of response for binary outcomes and will be implicitly handled by MMRM for continuous endpoints. Participants with a missing Week 24 assessment (Part II baseline) will be excluded from the respective Part II analysis.

The hypotheses specified below will be tested for satralizumab superiority hierarchically according to the graphical testing approach by [Bretz and Xi 2019](#) (see [Figure 1](#)) at an overall of two-sided significance level [REDACTED] (using $\alpha=0.05$) to control the overall type I error rate.

Figure 1 Graphical Alpha Control

α here represents two-sided 0.05; [REDACTED] for Part 2



Alpha will be adjusted by the number of iDMC reviews in Part I and Part II

[REDACTED]
[REDACTED] A fixed sequencing test approach (as specified below) will be used to control the overall type I error rate for the hypotheses in Part II. The hypotheses 1-5 below will address the question of maintenance of satralizumab treatment in Part II.

1. The proportion of participants maintaining proptosis response will be calculated for each (randomized) treatment group in Part II and compared across arms. Maintaining response is defined as participants who were classified as responders at Week 24 and who continued to meet the response definition at Week 48. The analysis will be done on the Part II FAS, restricted to active TED participants that were randomized and responded to satralizumab in Part I
2. The change in proptosis from Week 24 to 48 will be calculated and compared across (randomized) arms using MMRM. The analysis will be done on the Part II FAS restricted to active TED participants that were randomized and responded to satralizumab in Part I
3. The change in CAS from Week 24 to 48 will be calculated and compared across (randomized) arms using MMRM. The analysis will be done on the Part II FAS restricted to active TED participants that were randomized and responded to satralizumab in Part I
4. The proportion of participants with a ≥ 6 -point improvement in the Visual Functioning sub-scales) score of the GO-QoL from Week 24 at Week 48 will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS, restricted to participants with a visual functioning sub-scale of ≤ 94 at Week 24, and restricted to active TED participants that were randomized and responded to satralizumab in Part I
5. The proportion of participants with a ≥ 6 -point improvement in the Appearance sub-scale score of the GO-QoL from Week 24 at Week 48 will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS, restricted to participants with an appearance sub-scale of ≤ 94 at Week 24, and restricted to active TED participants that were randomized and responded to satralizumab in Part I
6. The proportion of participants maintaining response will be calculated for each (randomized) treatment group in Part II and compared across arms. Maintaining response is defined as participants who were classified as responders at Week 24 and who continued to meet the response definition at Week 48. The analysis will be done on the Part II FAS, (active TED participants).
7. The change in proptosis from Week 24 to 48 will be calculated and compared across (randomized) arms using MMRM. The analysis will be done on the Part II FAS (active TED participants)
8. The change in CAS from Week 24 to 48 will be calculated and compared across (randomized) arms using MMRM. The analysis will be done on the Part II FAS (active TED participants)

9. The proportion of participants with a ≥ 6 -point improvement in the Visual Functioning sub-scales) score of the GO-QoL from Week 24 at Week 48 will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS, restricted to participants with a visual functioning sub-scale of ≤ 94 at Week 24 (active TED participants)
10. The proportion of participants with a ≥ 6 -point improvement in the Appearance sub-scale score of the GO-QoL from Week 24 at Week 48 will be calculated for each (randomized) treatment group in Part II and compared across arms. The analysis will be done on the Part II FAS, restricted to participants with an appearance sub-scale of ≤ 94 at Week 24 (active TED participants)

The first hypothesis will correspond to the primary endpoint for Part II. The estimand framework for this endpoint is as follows:

- Population: adult participants with moderate to severe active thyroid eye disease that were re-randomized to Part II and were randomized to satralizumab in Part I.
- Variable: proportion of participants maintaining response will be calculated for each (randomized) treatment group in Part II and compared across arms. Maintaining response is defined as participants who were classified as responders at Week 24 and who continued to meet the response definition at Week 48.
- Treatment (as randomized):
 - Experimental: satralizumab SC [REDACTED] mg, [REDACTED] mg, or [REDACTED] mg (for bodyweight [REDACTED] respectively), at Week 24, Week 26, and Week 28 (loading doses) and maintenance doses Q4W thereafter until Week 48. Seven doses in total during the 24-week Part II period
 - Control: placebo SC at Week 24, Week 26, and Week 28 (loading doses) and maintenance doses Q4W thereafter until Week 48. Seven doses in total during the 24-week Part II period
- Intercurrent events occurring after Week 24 and prior to Week 48 and handling strategies:
 - Early discontinuation of study treatment due to AEs or lack of efficacy: treatment policy strategy
 - Use of any prohibited systemic treatment or prohibited therapy for either eye (study eye or fellow eye): evaluated as loss of response (composite strategy)
 - Administration of rescue treatment: evaluated as loss of response (composite strategy)
 - Orbital or ocular surgery/procedure for TED (e.g., orbital decompression, eyelid or strabismus surgery) in either eye (study eye or fellow eye): evaluated as loss of response (composite strategy)
- Population-level summary: difference in proportions of proptosis maintenance at Week 48 between the satralizumab and placebo group, measured as the weighted mean of stratum specific risk differences utilizing Mantel-Haenszel stratum weights

The hierarchical testing order of the Part II endpoints specified above may be updated after the Week 24 primary analysis (Note: the Sponsor will remain masked to the treatment assignment in Part II). However, the update would be specified in an updated version of the SAP which would be finalized prior to the Week 48 analysis and unmasking of this part of the study.

4.7 SAFETY ANALYSES

Safety will be assessed through descriptive summaries of exposure to study treatment, adverse events, changes in laboratory test results, and changes in vital signs and ECGs.

Clinically significant laboratory abnormalities and clinically relevant vital sign changes will be reported as AEs and evaluated as part of the AE assessments.

Missing data for safety analyses will not be imputed, with the exception of the imputation of the onset dates of AEs.

At the primary analysis, safety summaries will be performed based on data in Part I, i.e. prior to the administration of the study treatment at the Week 24 visit or 4 weeks after the last dosing in Part I, whichever occurs first. Every safety endpoint measured after the administration at the Week 24 visit belongs to Part II.

Safety data for Part I will be evaluated on the SAS.

For the analysis of Part I, participants will be grouped according to the treatment they actually received in Part I (satralizumab or placebo).

At the Week 48 analysis, safety summaries will be performed based on data in Part I and II (i.e., up to 4 weeks after the last dosing in Part II).

Safety data for Part I and II will be evaluated on the Part II SAS.

For the analysis of Part II, participants will be analyzed according to the different possible treatment combinations they received in Part I and II (SAT->SAT, SAT->PLB, PLB->PLB, PLB->SAT). No distinction between responders and non-responders is made. Non-responders who receive satralizumab in Part II are counted as “satralizumab” in Part II. Participants that discontinue study treatment in Part I but continue in Part II for follow up off-drug are counted as “placebo” in Part II.

For other analyses in Part II, participants will be grouped into “satralizumab” and “placebo” and their time on satralizumab and placebo will be combined. The incidence density for AE occurring in that period will be evaluated.

For the analysis of the follow up period AEs from Week 48 (or earlier if the participant discontinued the study treatment early) up to Week 72 are considered (final analysis).

In case of a mapping conflict of analysis visits within a visit window, for ECG and vital sign assessments, the value closest to the target day will be used.

For laboratory, the worst value will be used.

In the case of a tie, the value with the latest datetime will be used for all safety endpoints.

4.7.1 Extent of Exposure

At the primary analysis, exposure to study treatment in Part I (number of study drug administrations, total dose, and duration of treatment) will be summarized by treatment group based on data before the Week 24 visit for the SAS (the dose given at Week 24 belongs to Part II). Duration of treatment is the time from first study drug administration (Satralizumab or Placebo) to the earlier of:

- Date of treatment or study discontinuation before the Week 24 visit
- Date of last dose of study drug or last dose hold before the Week 24 visit
- Date of death
- CCOD

The Part II exposure outputs will be grouped by all treatment combinations (SAT -> SAT, SAT -> PLB, PLB -> SAT, and PLB -> PLB) and only the Part II treatment will be summarized (non-cumulative).

The following measures of exposure of satralizumab/placebo will be summarized using descriptive statistics for Part I and Part II separately:

- Treatment duration (weeks)
- Number of doses received (loading dose counts separately)
- Total cumulative dose (mg)
- Dosing regimen received at baseline (██████████ mg)
- Proportion of participants with at least one dose change'

4.7.2 Adverse Events

All verbatim adverse event terms will be mapped to the latest Medical Dictionary for Regulatory Activities (MedDRA) thesaurus terms, and adverse event severity will be graded according to National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE) v5.0. All adverse events, serious adverse events, adverse events leading to death, adverse events of special interest, and adverse events leading to study treatment discontinuation that occur on or after the first dose of study treatment (i.e., treatment-emergent adverse events) will be summarized by system organ class (SOC) and preferred term (PT), and severity grade. For events of varying severity, the highest grade per participant will be used in the summaries. Deaths and cause of death will be summarized.

Unless explicitly mentioned, only treatment-emergent adverse events (TEAE) will be included in the analyses. The term adverse event and treatment emergent adverse event will be used synonymously. Also, unless specified otherwise, all analyses of adverse events will be done on the Part I SAS (primary analysis) and Part II SAS (Week 48 and Week 72 analysis).

4.7.2.1 Definitions

Treatment-Emergent Adverse Event:

Adverse events (AE) with an observed or imputed date of onset on or after the start date of first study drug administration. If the onset date of the AE is prior to the day of first dose, the AE will be considered treatment-emergent only if the most extreme intensity is greater than the initial intensity (i.e., the intensity for a given AE increases and its end date is on or after the date of the first dose). An AE with a completely missing, non-imputed start date will be assumed to be treatment emergent unless the AE has a complete, non-imputed end date that is prior to the date of the first dose.

Imputation of incomplete date: in order to evaluate if an AE is treatment-emergent, incomplete dates will be imputed as follows:

- If the start date is incomplete:
 - If the day is missing, impute to 01, or to the day of first dose of study drug (satralizumab or placebo) if the month and year is the same as the dosing date.
 - If the month is missing, impute to JAN, or to the month of first dose of study drug if the year is the same as the year of dosing.
 - If the date is completely missing, then the date will remain missing.
- If the end date is incomplete:
 - If the day is missing, impute to the last day of the month.
 - If the month is missing, impute to DEC.
 - If the date is completely missing, then the date will be missing.

If the imputed start date of an AE is post first dose of study drug, it is considered treatment emergent.

If the imputed start date of an AE is prior first dose of study drug, it is considered non-treatment emergent.

In case of a missing start date and an (imputed) end date prior to first dose of study drug, the AE is considered non-treatment emergent.

Otherwise, the AE is considered treatment emergent.

Reporting Period for AEs in Part I (Primary Analysis):

Only AEs occurring in Part I are considered (i.e., AE with an imputed start date-time occurring prior to the study drug administration up to the Week 24 visit [beginning of Part II]). For participants that discontinue the study drug early in Part I, all AEs up to 4 weeks after the last dose are reported in the summary tables.

Reporting Period for AEs in Part I & Part II (Week 48 Analysis):

All AEs up to 4 weeks after the last dose are reported in summary tables.

Reporting Period for AEs in Follow up (Final Analysis):

All AEs occurring from four weeks after the last dose up until the end of the reporting period (Week 72) are reported in summary tables.

For all reporting periods, the listings will contain all AEs until the end of the corresponding period.

Output Conventions

All analyses of AE data will be performed using the PTs unless otherwise specified.

For all summary tables, the AEs will be sorted by SOC (in decreasing order of overall incidence) and then by PT (in decreasing order of overall incidence). At each level of summarization (at least one event, SOC and PT), participants reporting more than one AE will be counted only once. For AEs by grade (intensity), the highest grade per participant will be reported. All AE summaries will be based on AEs only.

The following summary tables will be generated:

- Overview summary table of AEs, including: total number of participants with at least one AE, total number of AEs, total number of deaths, total number withdrawn from study due to an AE, and total number of participants with at least one:
 - AE with fatal outcome
 - SAE
 - SAE leading to withdrawal from treatment
 - SAE leading to dose modification/interruption
 - Related SAE
 - AE leading to withdrawal from treatment
 - AE leading to dose modification/interruption
 - Related AE
 - Related AE leading to withdrawal from treatment
 - AE of Grade ≥ 3
 - Infections (use SOC: infections and infestations)
 - Infections of Grade ≥ 3 (use SOC: infections and infestations)

- Serious infections (use SOC: infections and infestations and filter for serious)
- Potential opportunistic infections (use specific MedDRA basket)
- Injection reactions
- Injection reactions of Grade ≥ 3
- Serious injection reactions
- Adverse event of special interest (AESI) (see Protocol Section 8.3.8.1): Cases of potential drug-induced liver injury, suspected transmission of an infectious agent by a study treatment
- Ophthalmic AEs of interest: Corneal Ulceration, dysthyroid optic neuropathy, and choroidal folds (use LLT “choroidal folds”).
- AEs, by SOC and PT
- Serious AEs, by SOC and PT
- AEs, by SOC and PT and intensity (grade)
- AEs leading to study drug discontinuation by SOC and PT
- Infections and serious infections by SOC and PT (using the Roche MedDRA basket)

All AEs, serious adverse events (SAE), AEs leading to study drug discontinuation, Adverse Events of Dysthyroid Optic Neuropathy, Adverse Events of Corneal Ulceration, anaphylactic cases, deaths and pregnancies occurring during the study will be reported as listings respectively. For Dysthyroid Optic Neuropathy, Adverse Events of Corneal Ulceration, the verbatim term will be included. AE leading to discontinuation will be added as a column in the listings.

All aforementioned tables and listings will be produced for Part I (primary analysis) and Part I+II (Week 48 analysis) respectively.

In addition, the overview summary table will be repeated for the subpopulation of active and inactive participants respectively and the all treated SAS for Part I and Part I+II respectively.

The overview summary table, an all-AE listing and the AEs by SOC and PT table will be repeated for the SAS Part I+II population for the AE-follow up period.

4.7.2.2 Laboratory Data

Laboratory data will be evaluated by treatment group in Part I and treatment group combination in Part I+II using summary statistics and displayed by time, with grades identified where appropriate. Hy's law for ALT/AST/Bilirubin will be added as well. Additionally, a shift table of selected laboratory tests will be used to summarize the baseline and maximum postbaseline severity grade. The adverse event reporting periods will be applied.

4.7.2.3 Vital Signs

Vital signs (weight, pulse rate, respiratory rate, blood pressure, and temperature) and vital signs change from baseline will be evaluated by treatment group in Part I and treatment group combination in Part I+II using summary statistics, and will be displayed by time. The adverse event reporting periods will be applied.

4.7.2.4 ECGs

ECG data and ECG change from baseline will be evaluated by treatment group in Part I and treatment group combination in Part I+II using summary statistics, and will be displayed by time. The adverse event reporting periods will be applied.

4.8 OTHER ANALYSES

4.8.1 Summaries of Conduct of Study

Enrollment, study treatment administration, and discontinuation from the study will be summarized by treatment arm in Part I and by treatment group combination in Part I+ II. The reasons for study treatment discontinuation will also be tabulated. Major protocol deviations, including major deviations with regard to the inclusion and exclusion criteria, will be summarized by treatment arm. The analysis will be done on all enrolled participants.

4.8.2 Summaries of Demographics and Baseline Characteristics

Demographics and baseline characteristics (including age, sex, race/ethnicity, region and iris pigmentation) will be summarized by treatment arm. Baseline data are the last data obtained prior to initiation of study treatment (satralizumab or placebo) on Study Day 1. Descriptive statistics (mean, SD, median, and range) will be presented for continuous variables and counts and percentages will be presented for categorical variables. The analysis will be done for Part I on the Part I FAS and for Part II on the Part II FAS.

4.8.3 Pharmacokinetic Analyses

The pharmacokinetic (PK) analysis population consists of all participants in the safety analysis set with at least one valid post-dose concentration result with a dosing record and sampling time. The trial will evaluate the PK characteristics of satralizumab treatment over 24 weeks in Part I of the study by summary statistics. PK samples will also be taken during Part II of the study, to further characterize the pharmacokinetics of satralizumab in TED over longer-term treatment. The serum concentration at each sampling timepoint will be described with means and SD of C_{trough} irrespective of whether participants receive rescue therapy, miss a dose, or if study drug administration is delayed, or if they withdraw from treatment before data collection at Week 24. Individual and mean serum-concentration-versus-time curves will be plotted. Non-linear mixed effects analysis will be performed to analyze the satralizumab concentration-time data collected in the trial. The model to be used was previously developed on the basis of PK data from adult healthy volunteers and adult and adolescent participants with NMOSD. Further model development may be undertaken if needed in order to achieve a

satisfactory description of the data, and the data from this study may be pooled with data from other studies with satralizumab. Population and individual PK and exposure parameters will be generated based on the model. Covariate analysis, including demographic factors and anti-drug antibody (ADA) status, will also be performed. Both the satralizumab concentration data and the results of the popPK analysis will be reported separately from the Clinical Study Report (CSR).

Further details are described in a separate Analysis Plan.

4.8.4 Immunogenicity Analyses

The immunogenicity analysis population will consist of all participants treated with satralizumab with at least one ADA assessment.

The number and proportion of ADA-positive participants and ADA-negative participants at baseline (baseline prevalence) and after drug administration (post-baseline incidence) will be summarized. When determining post-baseline incidence, participants are considered to be ADA positive if they are ADA negative or have missing data at baseline but develop an ADA response following study drug exposure (treatment-induced ADA response), or if they are ADA positive at baseline and the titer of one or more postbaseline samples is at least 0.60-titer unit greater than the titer of the baseline sample (treatment-enhanced ADA response). Participants are considered to be ADA negative if they are ADA negative or have missing data at baseline and all postbaseline samples are negative, or if they are ADA positive at baseline but do not have any postbaseline samples with a titer that is at least 0.60-titer unit greater than the titer of the baseline sample (treatment unaffected).

The percentage of participants who have positive or negative ADA results for satralizumab will be tabulated. Selected safety outputs will be summarized by anti-satralizumab antibody (i.e., satralizumab ADA) status.

The respective analyses will be done for Part I and then Part I and II combined.

4.8.4.1 Pharmacodynamic Analyses

Serum IL-6 and sIL-6R levels will be summarized by treatment group and timepoint descriptively, as appropriate.

4.9 INTERIM ANALYSES

4.9.1 Safety Review

Safety data will be reviewed regularly in an ongoing manner by the iDMC. [REDACTED]

[REDACTED]

[REDACTED]

5. SUPPORTING DOCUMENTATION

This section is not applicable, since there is no additional supporting document.

6. REFERENCES

- Allen RC, Bradley EA, Fante, RG, et al. A Perspective on the Current Role of Teprotumumab in Treatment of Thyroid Eye Disease. *Ophthalmology*. 2021;128(8):1125-8.
- Bahn RS. Graves' ophthalmopathy. *N Engl J Med*. 2010;362(8):726-38.
- Bahn RS. Autoimmunity and Graves' disease. *Clin Pharmacol Ther*. 2012;91(4):577-9.
- Barrio-Barrio J, Sabater AL, Bonet-Farriol E, et al. Graves' Ophthalmopathy: VISA versus EUGOGO Classification, Assessment, and Management. *J Ophthalmol*. 2015;2015:249125.
- Bartley GB. The epidemiologic characteristics and clinical course of ophthalmopathy associated with autoimmune thyroid disease in Olmsted County, Minnesota. *Trans Am Ophthalmol Soc*. 1994;92:477-588.
- Bretz F, Xi D. Commentary on "Statistics at FDA: Reflections on the Past Six Years", *Statistics in Biopharmaceutical Research*. 2019;11(1): 20-5.
- Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4), 417-451.
- Dik WA, Virakul S, van Steensel L. Current perspectives on the role of orbital fibroblasts in the pathogenesis of Graves' ophthalmopathy. *Exp Eye Res*. 2016;142:83-91
- Douglas RS, Kossler AL, Abrams J, et al. Expert Consensus on the Use of Teprotumumab for the Management of Thyroid Eye Disease Using a Modified-Delphi Approach. *J Neuroophthalmol*. 2022a;42(3):334-9.
- Ezra DG, Krell J, Rose GE, et al. Transcriptome-level microarray expression profiling implicates IGF-1 and Wnt signalling dysregulation in the pathogenesis of thyroid-associated orbitopathy. *J Clin Pathol*. 2012;65(7):608-13.
- Gerding MN, Terwee CB, Dekker FW, et al. Quality of life in patients with Graves' ophthalmopathy is markedly decreased: measurement by the medical outcomes study instrument. *Thyroid*. 1997;7(6):885-9.
- Gupta S, Douglas RS. The pathophysiology of thyroid eye disease: implications for immunotherapy. *Curr Opin Ophthalmol* 2011;22:385-90.
- ICH International Council for Harmonisation. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1) [resource on the Internet]. November 201]. Available at: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.
- Kahaly GJ, Hardt J, Petrak F, et al. Psychosocial factors in subjects with thyroid-associated ophthalmopathy. *Thyroid*. 2002;12(3):237-9.
- Khong JJ, McNab AA, Ebeling PR, et al. Pathogenesis of thyroid eye disease: review and update on molecular mechanisms. *Br J Ophthalmol*. 2016;100(1):142-50.

- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719-748.
- Patel A, Yang H, Douglas RS. A new era in the treatment of thyroid eye disease. *Am J Ophthalmol*. 2019;208:281–8.
- Rubin D.(1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.
- Sato, T. (1989). A method for estimating the variance of the Mantel-Haenszel risk difference. *Biometrics*, 45(4), 1173-1176.
- van Buuren S. 2008. "Flexible Imputation of Missing Data, Second Edition". (Chapman & Hall/CRC Interdisciplinary Statistics Series.)
- von Hippel, P. T., & Bartlett, J. W. (2021). Maximum likelihood multiple imputation: Faster imputations and consistent standard errors without posterior draws. *Statistical Science*, 36(3), 400-420. doi:10.1214/20-STS804
- Wang, N., & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4), 935-948. doi:10.1093/biomet/85.4.935
- Weetman AP, Cohen S, Gatter KC, et al. Immunohistochemical analysis of the retrobulbar tissues in Graves' ophthalmopathy. *Clin Exp Immunol*. 1989;75(2):222-7.
- Wiersinga WM, Bartalena L. Epidemiology and prevention of Graves' ophthalmopathy. *Thyroid*. 2002;12(10):855-60.

Signature Page for Statistical Analysis Plan for TED CSR
System identifier: RIM-CLIN-521324

Approval Task	 Company Signatory 25-Jul-2025 23:56:32 GMT+0000
---------------	---