J2G-MC-JZJC PRO Statistical Analysis Plan Version 2

LIBRETTO-431: A Multicenter, Randomized, Open-Label, Phase 3 Trial Comparing Selpercatinib to Platinum-Based and Pemetrexed Therapy with or without Pembrolizumab as Initial Treatment of Advanced or Metastatic RET Fusion-Positive Non-Small Cell Lung Cancer

NCT04194944

Approval Date: 27-Apr-2023

Patient-Reported Outcome Statistical Analysis Plan for Clinical Study J2G-MC-JZJC

LIBRETTO-431: A Multicenter, Randomized, Open-Label, Phase 3 Trial Comparing Selpercatinib to Platinum-Based and Pemetrexed Therapy with or without Pembrolizumab as Initial Treatment of Advanced or Metastatic RET Fusion-Positive Non-Small Cell Lung Cancer

LY3527723 (Selpercatinib; LOXO-292)

RET Fusion-Positive Non-Small Cell Lung Cancer

This is a multicenter, randomized, open-label, phase 3 trial comparing selpercatinib to platinum-based and pemetrexed therapy with or without pembrolizumab as initial treatment of advanced or metastatic ret fusion-positive non-small cell lung cancer

Eli Lilly and Company
Indianapolis, Indiana, USA 46285

Protocol J2G-MC-JZJC
Phase 3

Approved on 27 Apr 2023 GMT

Table of Contents

Version history

SAP Version History Summary

This is the second version. Updates from version 1_0 include:

- Incorporation of PRO secondary endpoint analysis, which was originally planned to be part of the main clinical SAP.
- Addition of meaningful change thresholds for NSCLC-SAQ derived during planned early data cut.
- Inclusion of descriptive summaries (tables and figures) for all 13 PRO-CTCAE items
- Addition of ITT-pembrolizumab population to the missing data analysis

1. Introduction

This statistical analysis plan (SAP) covers the secondary/tertiary and exploratory patient-reported outcomes (PRO). The clinical SAP covers the analysis methods for the primary and other secondary objectives.

This PRO SAP will focus on the analysis of PRO endpoints and the methodology for establishing clinically meaningful change thresholds for the NSCLC-SAQ.

Objectives	Endpoints
Primary	
• Not applicable	• Not applicable
Secondary	
• To compare the efficacy of selpercatinib with the combination of platinum-based (carboplatin or cisplatin) and pemetrexed therapy, with or without pembrolizumab, in patients with advanced or metastatic RET fusion-positive NSCLC	• Time to deterioration in pulmonary symptoms: cough, chest pain, and dyspnea, as measured by the NSCLC-SAQ total score
Tertiary/Exploratory	
• To compare patient-reported tolerability outcomes, including symptomatic adverse events and overall side-effect bother between patients treated with selpercatinib versus the combination of platinum-based (carboplatin or cisplatin) and pemetrexed therapy, with or without pembrolizumab	• Symptomatic Adverse Events: PRO-CTCAE • Overall Side-effect Bother (FACT-GP5)
• To compare patient-reported Physical Functioning and health-related quality of life (HRQoL) between patients treated with selpercatinib versus the combination of platinum-based (carboplatin or cisplatin) and pemetrexed therapy, with or without pembrolizumab	• Physical Functioning: EORTC QLQ-C30 Physical Functioning Subscale (and EORTC IL19) • Other HRQoL Outcomes: EORTC QLQ-C30, NSCLC-SAQ symptoms, EQ-5D-5L

Primary estimand/coprimary estimand

Not applicable

Secondary estimand(s)

See section 4.4.1.2.

1.1. Study Design

Study J2G-MC-JZJC is a global, multicenter, randomized, open-label, controlled Phase 3 study comparing selpercatinib (Arm A) to platinum-based and pemetrexed therapy with or without pembrolizumab (Arm B) in patients with advanced or metastatic, *RET* fusion-positive nonsquamous NSCLC.

Patients will be randomized using the following stratification factors:

- Geography (East Asia vs non-East Asia)
- Brain metastases per investigator assessment (presence vs absence or unknown), and
- Investigator's choice of treatment with or without pembrolizumab. This decision must be determined at the time of randomization.

Patients will be randomized 2:1 between two treatment arms (Arm A: Arm B) and will be treated until disease progression or other discontinuation criteria are met (Protocol section 7).

- Arm A: treated with selpercatinib (160 mg BID continuously in 21-day cycles) or
- Arm B: treated with pemetrexed (500 mg/m^2 IV) every 3 weeks plus the investigator's discretion of the following treatments administered every 3 weeks:
    - 4 cycles of carboplatin (AUC 5, maximum dose 750 mg IV) or cisplatin (75 mg/m^2 IV)
    - With or without pembrolizumab (200 mg IV) up to 35 cycles

After the completion of 4 cycles of chemotherapy without progressive disease, patients randomly assigned to Arm B will receive maintenance therapy with pemetrexed (500 mg/m^2) with or without pembrolizumab (200 mg) every 3 weeks according to the decision made at the time of randomization. If applicable pembrolizumab will be discontinued after completion of approximately 2 years of treatment (35 cycles of pembrolizumab). Patients with the intent by the investigator to be treated without pembrolizumab will be restricted to a maximum of 20% of the total sample size. Treatment will continue until radiographic disease progression confirmed by blinded independent central review (BICR), unacceptable toxicity, withdrawal of consent, or death.

Patients randomly assigned to Arm B who discontinue treatment for radiographic disease progression that is confirmed by BICR will be allowed cross over to selpercatinib (Arm A) if they meet the eligibility criteria for crossover. Crossover treatment will be optional at the discretion of the investigator.

The primary endpoint is PFS per RECIST 1.1. by BICR in the ITT pembrolizumab population and the ITT population. The ITT-pembrolizumab population which is defined as all randomized participants with investigator's intent to treatment with pembrolizumab if randomized to Arm B.

The planned sample size is 200 patients (140 PFS events) in the ITT-pembrolizumab population and 250 patients overall in the main ITT population.

2. Statistical Hypotheses

Not applicable.

2.1. Multiplicity Adjustment

There will be no adjustment for multiple testing for any of the analyses of the PRO endpoints.

3. Analysis Sets

For the purposes of analysis, the following analysis sets are defined (*Table 1*):

Table 1. Analysis sets

Participant Analysis Set	Description
Safety	All randomized patients who take at least 1 dose (including a partial dose) of study treatment. Analysis of safety data will be based on the actual treatment a patient received on the first study treatment administration regardless of which treatment they were randomized to receive ("as treated")
PRO Evaluable	All patients in the ITT population who are eligible for and consent to participation in the collection of PRO data.
Psychometric Evaluation	All participants from the ITT set with non-missing NSCLC-SAQ assessments at any timepoint
ITT/Enrolled	All randomized patients, even if a patient does not take the assigned treatment, does not receive the correct treatment, or otherwise does not follow the protocol. Patients will be analyzed according to the treatment arm they were assigned to regardless of what actual treatment they receive
ITT-pembrolizumab	Patients included in the ITT population who were stratified with the intent to receive pembrolizumab in the event of the control-arm assignment

4. Statistical Analyses

4.1. General Considerations

Statistical analysis specified in this PRO SAP will be the responsibility of Adelphi Values.

Continuous variables will be summarized using descriptive statistics (i.e., number of patients, mean, median, standard deviation, minimum, and maximum). Categorical variables will be summarized by frequency and its corresponding percentage.

All test of treatment effects will be conducted at a 2-sided alpha level of 0.05, unless otherwise stated, and all confidence intervals (CIs) will be given at a 2-sided 95% level.

The assumptions for each statistical method will be evaluated. If there is a violation of assumptions, alternative statistical methods may be used.

Additional exploratory analyses of the data will be conducted as deemed appropriate.

Statistical analysis will be performed using SAS software (SAS, version 9.2 or higher).

4.1.1. Definitions

Definitions of PRO variables will follow definitions specified in the clinical SAP.

- Baseline measurement: unless otherwise specified, the last non-missing measurement prior to the first dose of study drug for PRO analyses
- Duration: duration is calculated as
    - Duration (days): (end date – start date + 1)
    - Duration (weeks): (end date – start date + 1)/7
    - Duration (months): (end date – start date + 1)/30.4375
    - Duration (years): (end date – start date + 1)/365.25
- Study Day (PRO analyses): study day is calculated as assessment date – first dose date + 1 day if the assessment is done on or after the first dose day. If the assessment is done prior to the first dose day, study day will be calculated as assessment date – first dose day. Date of first dose is defined as Study Day 1.
- Time-to-event: the event or censoring time (days) is calculated as date of event/censoring – randomization date + 1

4.2. Participant Disposition

The number and proportion of patients in the ITT, ITT-pembrolizumab and PRO evaluable population will be summarized by treatment. Reasons for not providing PRO data, where reported, will be summarized (e.g., patient did not consent to PRO collection, or the instruments were not available in the correct language).

4.2.1. Available data and completion rates

The number of expected patients at each time point is defined as: any patient from the PRO evaluable population is expected to have a completed PRO assessment if they were eligible for PRO assessment and the nominal date for that visit is before both the end of study/data cut-off

date (DCO) date and before the date last known alive (i.e., the number of patients randomized minus patients who died, or terminated the study due to any reason at or before that visit).

Available and completion rates will be reported by the relevant schedule of assessment (SOA) for each PRO measure. The available data rate will be summarized by time point, defined as the proportion of patients who completed the questionnaire at that time point using the number of patients in the PRO evaluable population as denominator (fixed denominator).

The completion rate will be summarized by time point, defined as the proportion of patients who completed the questionnaire at that time point using the number of patients expected to have an assessment at the respective time using point as the denominator (variable denominator). These tables will be presented by treatment arm.

4.2.2. Missing data

For analysis of missing data, the ITT population and ITT-pembrolizumab population will be used, and analysis will be conducted for the NSCLC-SAQ, EORTC QLQ-C30, and IL19.

4.2.2.1. Missing data patterns

The number of patients in each treatment arm and overall will be summarized for each on-treatment visit where at least one patient had their last non-missing assessment (relating to a unique missing data pattern), providing all possible missing data patterns.

The number and proportion of patients will be summarized for all data points described in the categories below where 'X' denotes a visit with a PRO value and 'O' denotes a visit with a missing PRO value (e.g., XXOXXOXX would be a mixed pattern with PRO values present at baseline and all but two cycles).

The proportion of patients with each missing data pattern will also be summarized by treatment arm and overall. Examples based on the first five time points showing patterns of drop-outwill be defined as shown below:

- No PRO data (no assessment at any time point – OOOOO);
- Monotone pattern (all assessments missing after a certain visit and all non-missing assessments before first missing visit – XXXXO);
- Intermittent missing (has non-missing assessment at last on-treatment visit with one or more missing assessments prior to last on-treatment visit – XXXOX);
- Mixed missing (last assessment at last on-treatment visit is missing and has one or more missing assessments at earlier timepoints – XXOXO)
- Complete (all on-treatment assessments are non-missing XXXXX).

4.2.2.2. Reasons for missing data

The reasons for missing PRO data at baseline and at each assessment time point with planned PRO data collection will be presented. The number and proportion of patients in PRO evaluable population with missing PRO data will be presented, overall and by treatment arm. The reasons for missing data at each on-treatment time point will be presented using the following categories, where reasons at that time point have been collected:

- Expected: Non-missing
- Expected: Missing – no valid PRO assessment
- Not Expected: Missing due to:
    - Study site failed to administer PRO
    - Participant was unable to complete
    - Technical inability
    - Translation not available
    - Participant refused
    - Participant missed appointment
    - Unknown
    - Other
    - Death
    - Discontinued

4.2.2.3. Timing of assessment plots

Using the ITT population and ITT-pembrolizumab population, scatterplots, and assessments by treatment arm will be used to visualize the variation in assessment timing.

A scatterplot will be produced, for each treatment arm, with each subject on the y-axis and time since randomization in days on the x-axis. A point will be plotted for each of a subject's assessments. Markers will be assigned to each point according to whether they were assigned to a visit (e.g., baseline, on-treatment, follow-up) or were discarded (e.g., in the event of two assessments within the same time window). This will be used to compare the comparability of the timing of PRO assessments between the treatment arms.

4.2.2.4. Relationship of missing data with observed PRO scores

Using the ITT population and ITT-pembrolizumab population, graphs will be produced for the NSCLC-SAQ total score and EORTC QLQ-C30/IL19 physical functioning score by treatment arm showing the mean PRO score over time stratified by the time of last on-treatment assessment (dropout) along with the number of patients in each of these dropout groups. The aim is to visualize the association between time of dropout and observed PRO scores and whether this is different across the treatment arms (i.e., patients with more severe pain at baseline drop out at an earlier time point). If groups are small ($n<10$) or there are many patterns (e.g., for the PROs with weekly assessments), neighboring groups may be combined for ease of interpretation

4.3. Primary Endpoint Analysis

Not applicable.

4.4. Secondary Endpoint Analysis

4.4.1. PRO measures

The time to deterioration (TTD) in NSCLC-SAQ total score is evaluated as a secondary endpoint. Clinically meaningful thresholds for the NSCLC-SAQ were determined in the meaningful change analysis (methods are provided in Section 4.7.3) and are reported in Section 4.4.1.1.

4.4.1.1. NSCLC-SAQ

The NSCLC-SAQ (McCarrier et al., 2016, Williams et al., 2022) consists of seven items assessing five NSCLC symptom concepts: cough, pain, dyspnea, fatigue, and poor appetite. All items have a recall period of past 7 days and a 5-point response scale ranging from 0: "Not at All" to 4: "Very severe" or from 0: "Never" to 4: "Always" to measure attributes of symptom intensity or frequency, respectively. Two pain items form a single "Pain" domain representing the most severe response of the two items, and two fatigue items form a single "Fatigue" domain by calculating their mean. The NSCLC-SAQ total score is computed as the sum of the 5 domains and ranges between 0 and 20. Higher scores indicate more severe symptomatology.

Estimates of the within group change (minimal important change [MIC]), between group difference (minimal important difference [MID]), and meaningful within patient change (MWPC) thresholds for deterioration and improvement are provided in Table 2. These were derived using an early data cut (see Section 4.7.3) and added to Version 2_0 of this PRO SAP. The MWPC estimate to be used for the secondary endpoint of TTD for NSCLC-SAQ total score is 2 points.

Table 2: NSCLC-SAQ Meaningful Change Threshold Estimates based on analyses of early data

NSCLC-SAQ	Deterioration			Improvement		
Threshold	**MIC**	**MID**	**MWPC**	**MIC**	**MID**	**MWPC**
Total score	0.5	2.0	2.0	-3.5	-2.0	-2.5
Cough	1.0	0.5	1.0	-1.5	-0.5	-1.0
Pain	0.5	1.0	1.0	-1.0	-1.0	-1.0
Dyspnea	0.5	1.0	1.0	-1.0	-1.0	-1.0
Fatigue	0.5	0.5	0.5	-1.0	-1.0	-1.0
Appetite	1.0	1.0	1.0	-1.5	-1.0	-1.0

4.4.1.2. Time to deterioration in NSCLC-SAQ total score estimand

The estimand for the secondary PRO objective of the time to confirmed deterioration (TTD) in NSCLC-SAQ total score is described by the following attributes:

- Population: adult participants with advanced or metastatic *RET* fusion-positive NSCLC with no previous systemic therapy for metastatic disease.
    - For the comparison between Arm A (selpercatinib) and Arm B (platinum-based and pemetrexed therapy with or without pembrolizumab), the analysis population is based on approximately 250 patients randomized to both arms (ITT population).
    - For the comparison between Arm A (selpercatinib) and Arm B (platinum-based and pemetrexed therapy with pembrolizumab), the analysis population is based on approximately 200 patients randomized to both arms (ITT-pembrolizumab population).
- Endpoint: time to confirmed deterioration in NSCLC-SAQ item
    - The time from the date of randomization to the date of the first increase (≥ 2 points) confirmed at the next subsequent assessment.
- Treatment condition: the randomized study interventions (Arms A and B) will be administered until disease progression confirmed by BICR, unacceptable toxicity, withdrawal of consent or death.
- Intercurrent-event strategies (IES):
    - Death – patients with no increase in NSCLC-SAQ score prior to death will be censored at the week of their last assessment prior to death.
    - Disease progression – patients with no increase in NSCLC-SAQ score prior to disease progression will be censored at the week of their last assessment prior to disease progression.
    - Treatment discontinuation due to toxicity/physician decision/subject withdrawal – patients who discontinue study treatment with no increase in NSCLC-SAQ score prior to treatment discontinuation will be censored at the week of their last assessment prior to treatment discontinuation.
- Population-level summary measures:
    - Median time to deterioration in NSCLC-SAQ total score with 95% CI for arms A and B.
    - P-value of stratified log rank test for Arm A vs Arm B (test stratified for the randomization factors).
    - Hazard ratio (HR) with 95% CI for Arm A vs Arm B, from a stratified Cox proportional hazards model stratified by the randomization factors.

Rationale for IES: The interest lies in the treatment effect while on study treatment without the confounding effect of patients on Arm B crossing over to Arm A upon disease progression.

4.4.1.3. Main analytical approach

The Kaplan-Meier method will be used to estimate the TTD in NSCLC-SAQ total score for each treatment arm. Plots will be presented for each of the two treatment arms and median TTD will be reported for each group with the 95% CI. If the median TTD is not reached for both arms, then the 25% percentile estimate will be reported with the 95% CI. Treatment arms will be compared using a stratified log-rank test (stratified by the randomization factors) and the p-values comparing Arms A and B will be reported. The treatment effect between Arms A and B will be estimated using a stratified Cox proportional hazards model adjusted for treatment and stratified by the randomization factors, and the hazard ratio (HR) will be reported together with the 95% CI. Details of the censoring are reported in Table 3 below.

All TTD analyses will be performed in the ITT-pembrolizumab and ITT populations.

Table 3: TTD Censoring

Situation	Event/Censor	Date of Event or Censor
Deterioration (Increase in PRO score)	Event	Date of first visit with an increase of 2 points in the PRO score confirmed at next subsequent visit
No deterioration while on treatment	Censored	Date of last PRO assessment
Death	Censored	Date of last PRO assessment prior to death, or date of randomization (whichever is later)
Disease progression	Censored	Date of last PRO assessment prior to progression, or date of randomization (whichever is later)
Treatment discontinuation	Censored	Date of last PRO assessment prior to treatment discontinuation, or date of randomization (whichever is later)
Missing baseline and/or post-baseline data	Censored	Date of randomization Cycle 1 Day 1 (C1D1) if no PRO data. Day 2 if baseline PRO data but no post-baseline.

4.4.1.4. Sensitivity analyses

A sensitivity analysis will include death as an event rather than a censoring event for TTD. In this analysis patients will be counted as having event if they have a deterioration at two consecutive time points or die (before deterioration, or after the first deterioration but before the next assessment). An additional analysis will use the time to first deterioration (TTFD) defined

as the time from randomization to the date of the first increase in NSCLC-SAQ total score (≥ 2 points). The analyses specified in Section 4.4.1.3 will be repeated for these endpoints.

4.5. Tertiary/Exploratory Endpoints Analysis

The tertiary/exploratory PRO endpoints are physical functioning measured with EORTC QLQ-C30 (and IL19); other QLQ-C30 functioning and symptom subscales; NSCLC-SAQ symptoms; EQ-5D-5L; FACT-GP5 "overall burden of side effects"; and PRO-CTCAE items .

4.5.1.1. EORTC QLQ-C30 and IL19

The EORTC QLQ-C30 is measured at baseline (C1D1)) and every 3 weeks to short term follow-up (approximately 30 (+/- 7) days from the discontinuation of study treatment. The EORTC IL19, items 1-5 of the EORTC QLQ-C30 Physical Functioning scale, is measured on the weeks that the QLQ-C30 is not administered so physical functioning is completed every week.

The EORTC QLQ-C30 is a validated self-rating questionnaire including 30 items. It is composed of multi-item and single-item scales. The QLQ-C30 includes:

- One scale for Global Health Status / Quality of Life (QOL)
- Five Functional Scales:
    - Physical Functioning;
    - Role Functioning;
    - Emotional Functioning;
    - Cognitive Functioning;
    - Social Functioning.
- Nine Symptom Scales:
    - Fatigue;
    - Nausea/Vomiting;
    - Pain;
    - Dyspnea;
    - Insomnia;
    - Appetite Loss;
    - Constipation;
    - Diarrhea;
    - Financial Difficulties.

Data will be scored according to the algorithm described in the EORTC QLQ-C30 scoring manual. All scales and single items are linearly transformed to 0-100 scales where:

- A high score for a symptom scale or item represents a high level of symptoms or problems.
- A high score for a functional scale represents a high or healthy level of functioning.
- A high score for the global health status/QoL represents high QoL.

Total scores will be calculated for the global HRQoL scale, for the functional scales and for each symptom scale/item, from the scored categorical scales as follows:

Global health status:

- Global health status/QoL: ((Q29+Q30)/2-1)/6*100

Functional scales:

- Physical functioning: (1-((Q1+Q2+Q3+Q4+Q5)/5-1)/3) *100
- Role functioning: (1-((Q6+Q7)/2-1)/3) *100
- Emotional functioning: (1-((Q21+Q22+Q23+Q24)/4-1)/3) *100
- Cognitive functioning: (1-((Q20+Q25)/2-1)/3) *100
- Social functioning: (1-((Q26+Q27)/2-1)/3) *100

Symptom scales/items:

- Fatigue: ((Q10+Q12+Q18)/3-1)/3*100
- Nausea and vomiting: ((Q14+Q15)/2-1)/3*100
- Pain: ((Q9+Q19)/2-1)/3*100
- Dyspnea: ((Q8-1)/3*100
- Insomnia: (Q11-1)/3*100
- Appetite loss: (Q13-1)/3*100
- Constipation: (Q16-1)/3*100
- Diarrhea: (Q17-1)/3*100
- Financial difficulties: (Q28-1)/3*100

The EORTC QLQ-C30 scoring manual recommends application of the half scale rule in consideration of missing data (Fayers et al., 2001). Following this approach, if fewer than half of the items on a given scale have been answered, the scale score is set to missing. If at least half of the items from the scale have been answered, the developers recommend calculation of the scale score using all of the items that were completed (ignoring any items with missing values). Missing values will then be imputed by assuming that the missing items have values equal to the average of those items which are present for any scale in which at least half the items are completed. For missing responses on any single-item measures, the score is set to missing. A questionnaire will be considered as received if at least one of the 15 scales is non-missing (after imputation).

Established guidelines from Cocks et al. will be used to help inform the interpretation of changes in scores over time and between group differences (see Table 4 and Table 5). Recently published guidelines, which are based on two EORTC-sponsored RCTs in advanced NSCLC (Koller et al., 2022), will be used to assess which end of the range of thresholds from Cocks et al is appropriate for this setting. In general, the smallest non-trivial threshold will be used from Cocks et al (1 point higher than the lower end of the range provided as the lowest point represents the upper bound of trivial differences) unless the Koller paper suggests a higher value would be more appropriate. Thresholds will not be reduced below the recommended smallest non-trivial difference. Details of responder definitions for response within an individual participant highlighting the range of possible change scores for each subscale are shown in Table 6. A responder definition could be considered the smallest amount of change possible on each scale with the next change increment as a sensitivity analysis (Cocks et al., 2011, Cocks et al., 2012) but note the level of individual change threshold has not been confirmed as meaningful using qualitative data.

Table 4: EORTC QLQ-C30 and Koller Guidelines for Group Changes Over Time

Subscale	Deterioration			Improvement		
	Cocks: Non-trivial *(range)	Koller	Consensus	Cocks: Non-trivial (range)	Koller	Consensus
Global health status/QoL	-6 (-10 to -5)	-5	-6	6 (5 to 8)	5	6
Physical functioning	-6 (-10 to -5)	-7 (-8 to -6)	-6	3 (2 to 7)	6	6
Role functioning	-8 (-14 to -7)	-9 (-11 to -7)	-8	7 (6 to 12)	9	9
Social functioning	-7 (-11 to -6)	-5	-7	4 (3 to 8)	6	6
Cognitive functioning	-2 (-7 to -1)	Not specified	-2	4 (3 to 7)	Not specified	4
Emotional functioning	-4 (-12 to -3)	Not specified	-4	7 (6 to 9)	Not specified	7
Nausea and vomiting	-6 (-11 to -5)	-13 (-14 to -13)	-13	4 (3 to 9)	Not specified	4
Dyspnea	-6 (-11 to -5)	Not specified	-6	3 (2 to 9)	Not specified	3
Fatigue	-6 (-10 to -5)	-9 (-10 to -8)	-9	5 (4 to 9)	6	6

Table 4: EORTC QLQ-C30 and Koller Guidelines for Group Changes Over Time

Subscale	Deterioration			Improvement		
	Cocks: Non-trivial *(range)	Koller	Consensus	Cocks: Non-trivial (range)	Koller	Consensus
Insomnia	-3 (-9 to -2)	Not specified	-3	6 (5 to 9)	Not specified	6
Pain	-4 (-11 to -3)	-12	-12	6 (5 to 9)	9 (8 to 10)	9
Constipation	-6 (-15 to -5)	-10	-10	5 (4 to 10)	13	13
Diarrhea	-6 (-15 to -5)	Not specified	-6	4 (3 to 11)	Not specified	4
Appetite loss	-3 (-14 to -2)	-11 (-15 to -8)	-11	8 (7 to 13)	8 (6 to 11)	8
Financial difficulties	-3 (-10 to -2)	Not specified	-3	4 (>3)	Not specified	4

* 1-point higher than the lower end of the range

Table 5: EORTC QLQ-C30 and Koller Guidelines for Between-Group Minimally Important Differences

Subscale*	Cocks: Small* (range)	Koller Deterioration*	Koller Improvement**	Consensus
Global health status/QoL	5 (4 to 10)	-4	4	5
Physical functioning	6 (5 to 14)	-4	5	6
Role functioning	7 (6 to 19)	-6 (-8 to -4)	7	7
Social functioning	6 (5 to 11)	-4	5	6
Cognitive functioning	4 (3 to 9)	Not specified	Not specified	4
Emotional Functioning	Not specified	Not specified	Not specified	Not specified
Nausea and vomiting	4 (3 to 8)	-8 (-9 to -7)	Not specified	8
Dyspnea	5 (4 to 9)	Not specified	Not specified	5
Fatigue	6 (5 to 13)	-6 (-6 to -5)	6	6
Insomnia	5 (4 to 13)	Not specified	Not specified	5
Pain	7 (6 to 13)	-9	9 (7 to 10)	9
Constipation	6 (5 to 13)	-9	13	13
Diarrhea	4 (3 to 7)	Not specified	Not specified	4
Financial difficulties	4 (3 to 10)	Not specified	Not specified	4
Appetite loss	6 (5 to 14)	-7 (-9 to -5)	10 (6 to 15)	10

*1-point higher than the lower end of the range

** Koller et al specify separate guidelines for deterioration and improvement for between-group differences. In practice these are normally combined as one threshold since for between-group differences the direction is not important. The consensus here takes into account both ranges provided by Koller.

Table 6: EORTC QLQ-C30 plausible responder definitions

EORTC QLQ-C30 Subscale	Number of Items in Subscale	Increments in individual participant changes scores	Responder definitions (minimal change) **	Direction of change relating to deterioration
GHS/QOL	Two*	8.3	≥5	Negative
Functional Domains				
Physical functioning	Five	6.7	≥5	Negative
Emotional functioning	Four	8.3	≥5	Negative

Table 6: EORTC QLQ-C30 plausible responder definitions

EORTC QLQ-C30 Subscale	Number of Items in Subscale	Increments in individual participant changes scores	Responder definitions (minimal change) **	Direction of change relating to deterioration
Role, cognitive, social functioning	Two	16.7	≥15	Negative
Symptom Scales				
Fatigue	Three	11.1	≥10	Positive
Nausea and vomiting, pain	Two	16.7	≥15	Positive
Dyspnea, Insomnia, appetite loss, constipation, diarrhea, financial difficulties	Single	33.3	≥30	Positive

GHS/QOL – Global Health Status/Quality of Life; *GHS/QOL two items, 7-point Likert scale. All other subscales have 4-point Likert scale; **Minimal change is moving 1 response category on the Likert scale on 1 item in the scale (applicable to all subscales), round to nearest 1 decimal place.

4.5.1.2. PRO-CTCAE

Thirteen items from the PRO-CTCAE are being collected at baseline and every week. The two key items of interest are: "Decreased Appetite" and "Fatigue, Tiredness, or Lack of Energy" at its worst. The patient is asked to describe their experiences over the last seven days for the frequency (never, rarely, occasionally, frequently, almost constantly), severity of the symptom (none, mild, moderate, severe, and very severe), and how much it interfered with their usual or daily activities (not at all, little bit, somewhat, quite a bit, and very much).

4.5.1.3. FACT-GP5

This is a single item from the Functional Assessment of Cancer Therapy-Side Effects general scale to assess the overall burden of the items reported via the PRO-CTCAE measured on a weekly basis. The GP5 item "I am bothered by side effects of treatment" is an ordered categorical variable with responses of 0 "Not at all," 1 "A little bit," 2 "Somewhat," 3 "Quite a bit," and 4 "Very much".

4.5.1.4. EQ-5D-5L

Overall health status will be assessed using the five-level version of the EuroQoL Group's self-reported health status measure EQ 5D (EQ-5D-5L) which is measured at baseline (day 1, cycle 1) and every three weeks. This has two components, the EQ 5D descriptive system and the EQ 5D VAS. The EQ 5D 5L descriptive system comprises the following five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has five levels: (1) no problems, (2) slight problems, (3) moderate problems, (4) severe problems, (5) extreme problems. A unique health state is defined by combining one level from each of the five dimensions. EQ 5D 5L health states, defined by the EQ 5D 5L descriptive system, may be converted into a single summary index by applying a formula that essentially attaches values (also called weights) to each of the levels in each dimension. The index can be calculated by

deducting the appropriate weights from 1, the value for full health (i.e., state 11111). Information in this format is useful, for example, in cost utility analysis. Value sets have been derived for EQ 5D 5L in several countries using the EQ 5D VAS valuation technique or the time trade-off valuation technique. The United Kingdom (UK) Measurement and Valuation of Health study value set is generally considered the base case scoring function for the purposes of publication. Therefore, all EQ 5D utility index scores will be based on UK values.

The EQ 5D 5L VAS records the subject's self-rated health state on a 100-point vertical VAS ranging from 0="Worst imaginable health state" to 100="Best imaginable health state." Results for EQ-5D will be descriptive only, no statistical analysis will be performed.

4.5.2. Descriptive summaries

The EORTC QLQ-C30/IL19 physical functioning subscale and other EORTC QLQ-C30 subscales, EQ-5D-5L utility index and VAS, and NSCLC-SAQ symptoms and total score will be treated as continuous variables. The raw scores and change from baseline scores will be tabulated by presenting the number of subjects with non-missing data (n), mean, standard deviation, median, 25th and 75th percentiles, minimum, and maximum. These data will also be provided graphically (means +/- one standard error for raw scores and for change from baseline scores) and labeled with Ns per time point. Reference lines showing the minimally important group changes will be shown on the change from baseline figures using the values in the previous tables.

The proportion of patients improved/stable/deteriorated, according to the responder definitions provided above, at each time point will be summarized for all EORTC QLQ-C30 subscales, NSCLC-SAQ total score and symptoms. These data will initially be presented using the ITT population as the denominator with missing data as a category. Additionally, these data will be presented using only the non-missing responses at each visit, so that the proportions of patients improved/stable/deteriorated at each time point total 100% to assess if the proportions remain consistent over time. PRO responder status by time point will be presented in stacked bar charts. PRO responder status by time point will be presented in stacked bar charts, these bar charts will only be produced for visits with at least 10 patients in both treatment arms.

For ordered categorical variables (PRO-CTCAE items and FACT-GP5) the number and proportion of patients reporting each response category at each time point will be summarized by treatment arm using the same methods as for the proportion of patients improved/stable/deteriorated. These data will also be presented in stacked bar charts. The FACT GP5 side effects of bother will also be presented as a dichotomous outcome (bother (somewhat/quite a bit/very much) /no bother (not at all/ a little) and plotted on bar charts at each time point by treatment arm.

All descriptive summary tables and plots will be presented for both the ITT and ITT-pembrolizumab populations.

4.5.3. Longitudinal analysis

4.5.3.1. Growth curve models

Changes over time in the NSCLC-SAQ total score, and EORTC QLQ-C30/IL19, physical functioning scores will be compared between treatment groups using a growth curve model. Mixed-effect growth curve models will be used to conceptualize time as a continuous variable using piecewise linear regression with a change point at Week 12 to allow for variation in rates of change across treatment arms. The change point at Week 12 has been chosen to align with the addition of maintenance therapy after four cycles in Arm B as the score trajectory may change as the treatment changes.

The growth curve models are built to include both a mean and covariance structure whilst adjusting for baseline score by fixing a common baseline across both arms. The mean model structure for the two treatment arms, below, show that the slope of the curve can vary between arms, whereas the baseline and time change point are set to be equal. The random effects will use an unstructured covariance structure allowing for the rate of change over time to vary between patients.

Arm B: $Y_{ij}(t) = \beta_0 + \beta_1 t_{ij} + \beta_3 t_{ij}^{[12]} + \varepsilon_{hij}$,

Arm A: $Y_{ij}(t) = \beta_0 + \beta_2 t_{ij} + \beta_4 t_{ij}^{[12]} + \varepsilon_{hij}$,

Where Y_{ij} represents the jth HRQL score for patient i, β_0 represents the shared intercept at Baseline, t represents time and $\beta_{1\text{-}4}$ represent estimated coefficients which act on time and form the shape of each arm's trajectory. The $t^{[12]}$ represents the point in time where the shape of the curve is expected to change

The randomization stratification factors will be included in the growth curve model as covariates independently and as an interaction term with time, time will be included as a continuous variable.

The overall post-randomization treatment differences between the treatment arms will be compared by simultaneously comparing the parameters of the two curves. The following null hypothesis is tested for superiority with a p value of <0.05; H_0: $\beta_1 = \beta_2$, $\beta_3 = \beta_4$.

A contrast test will be used to estimate and compared the mean score at week 12.(Fairclough, 2002) The contrast tests will be conducted by estimating the mean HRQL for the control arm (Arm B) following the estimated model;

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_3 t^{[12]}$, where at 12 weeks, t = 12 and $t^{[12]} = 0$.

After analysis of the contrast tests, it may be appropriate to perform model reduction and remove the change point, if no change in slope is observed between specific treatment periods, to create a more parsimonious model. If the proposed change point is not identified as significant then a general linear model will be implemented.

The growth curve analysis assumes that the missing data is MAR. Should the assumption of MAR be unreasonable sensitivity analyses be carried out to account for possible non-ignorable missing data.(Fairclough, 2002, Bell and Fairclough, 2014)

The following will be reported:

- Estimates of LS means overall and at week 12 for each treatment, together with the difference (Arm A – Arm B), 95% CI and p-value
- The area under the curve (AUC) up to week 12 for each treatment
- The rates of change (slope) from weeks 0 to 12, and > 12 weeks for each treatment

4.5.3.2. Mixed model for repeated measures

The remaining EORTC QLQ-C30 subscales for functioning and symptoms which are measured every 3 weeks, will be evaluated using a longitudinal mixed model for repeated measures (MMRM) based on the raw score at each on-treatment visit (baseline and post-baseline), for the ITT and ITT-pembrolizumab populations. The baseline score and score for all on-treatment visits will be included in the model. If the data become very sparse over time, the model will only be fitted for on-treatment visits with at least 10 patients in each treatment arm (to be confirmed based on review of the available data); any cut-off used will be noted in footnotes.

The specific statistical model to be used will be a constrained longitudinal data analysis (Coffman et al., 2016) (CLDA) model which assumes a common mean across treatment arms at baseline due to randomization and is based on a multivariate normal distribution for the response, so patients with missing baseline (but post-baseline) or post-baseline (but no baseline) are still included in the likelihood function for estimation and inference. The dependent variable in the models will be the raw PRO score at each on-treatment visit (baseline and post-baseline). The model will contain treatment arm, study visit (as a categorical variable), and the randomization stratification factors from the interactive response technology as fixed effects. An interaction between treatment and study visit will also be specified as a fixed effect. Study visit will be fitted as a repeated effect (repeated by subject). An unstructured covariance matrix will be used. If, after ensuring sufficient patients at each timepoint and in each level of the categorical dependent variables (Fairclough, 2010), the mixed model does not converge using an unstructured covariance structure, first a heterogeneous covariance structure will be used, then a compound symmetry covariance structure will be used in SAS PROC MIXED. Parameters will be estimated using restricted maximum likelihood with the Newton–Raphson algorithm and using the Kenward–Roger method for calculating the denominator degrees of freedom. The covariance structure used will be noted in the footnote along with any cut-off point used, if applied (e.g., time points with at least 10 patients per treatment arm).

The number of patients contributing to the overall model and at each on-treatment time point (only those visits with more patients than agreed cut-off) will be presented together with the following estimates from the CLDA model.

Change from baseline overall estimate of:

- LS mean and 95% CI for each treatment
- Difference in LS mean (Arm A minus Arm B) between treatment arms and 95% CI for the difference
- P-value for the difference in LS means between treatment arms

Change from baseline results at each visit, estimates of:

- LS mean and 95% CI for each treatment

- P-value for the difference in LS means between treatment arms

The estimated treatment difference with 95% CI for the change from baseline in PRO score from the CLDA model will be plotted by visit and overall.

4.5.4. Time to deterioration analysis

4.5.4.1. Time to deterioration in physical functioning estimand

The estimand for the exploratory PRO objective of the time to deterioration (TTD) in physical functioning is described by the following attributes:

- Population: adult participants with advanced or metastatic *RET* fusion-positive NSCLC with no previous systemic therapy for metastatic disease.
    - For the comparison between Arm A (selpercatinib) and Arm B (platinum-based and pemetrexed therapy with or without pembrolizumab), the analysis population is based on approximately 250 participants randomized to both arms (ITT population).
    - For the comparison between Arm A (selpercatinib) and Arm B (platinum-based and pemetrexed therapy with pembrolizumab), the analysis population is based on approximately 200 participants randomized to both arms (ITT-pembrolizumab population).
- Endpoint: time to confirmed deterioration in physical functioning measured using EORTC QLQ-C30 and IL19
    - The time from the date of randomization to the date of the first increase (≥ 5 points) confirmed at the next subsequent assessment.
- Treatment condition: the randomized study interventions (Arms A and B) will be administered until disease progression confirmed by BICR, unacceptable toxicity, withdrawal of consent or death.
- Intercurrent-event strategies (IES):
    - Death – patients with no increase in physical functioning score prior to death will be censored at the week of their last assessment prior to death.
    - Disease progression – patients with no increase in physical functioning score prior to disease progression will be censored at the week of their last assessment prior to disease progression.
    - Treatment discontinuation due to toxicity/physician decision/subject withdrawal – patients who discontinue study treatment with no increase in physical functioning score prior to discontinuation will be censored at the week of their last assessment prior to treatment discontinuation.
- Population-level summary measures:
    - Median time to deterioration in physical functioning with 95% CI for arms A and B.
    - P-value of stratified log rank test for Arm A vs Arm B (test stratified for the randomization factors).
    - Hazard ratio (HR) with 95% CI for Arm A vs Arm B, from a stratified Cox proportional hazards model stratified by the randomization factors.

Rationale for IES: The interest lies in the treatment effect while on study treatment without the confounding effect of patients on Arm B crossing over to Arm A upon disease progression.

4.5.4.2. Main analytical approach

The Kaplan-Meier method will be used to estimate the TTD for each treatment arm. Plots will be presented for each of the two treatment arms and median TTD will be reported for each group with the 95% confidence interval (CI) If the median TTD is not reached for both arms, then the 25% percentile estimate will be reported with the 95% CI. Treatment arms will be compared using a stratified log-rank test (stratified by the randomization factors) and the p-values comparing Arms A and B will be reported. The treatment effect between Arms A and B will be estimated using a stratified Cox proportional hazards model adjusted for treatment and stratified by the randomization factors, and the hazard ratio (HR) will be reported together with the 95% CI. Details of the censoring are reported in Table 7 below.

TTD analyses of the other EORTC QLQ-C30 subscales and NSCLC-SAQ symptoms will use the same methods. All TTD analyses will be performed in the ITT-pembrolizumab and ITT populations.

Table 7: TTD Censoring

Situation	Event/Censor	Date of Event or Censor
Deterioration (Increase in PRO score)	Event	Date of first visit with an increase in the PRO score (based on relevant threshold) confirmed at next subsequent visit
No deterioration while on treatment	Censored	Date of last PRO assessment
Death	Censored	Date of last PRO assessment prior to death, or date of randomization (whichever is later)
Disease progression	Censored	Date of last PRO assessment prior to progression, or date of randomization (whichever is later)
Treatment discontinuation	Censored	Date of last PRO assessment prior to treatment discontinuation, or date of randomization (whichever is later)
Missing baseline and/or post-baseline data	Censored	Date of randomization Cycle 1 Day 1 (C1D1) if no PRO data. Day 2 if baseline PRO data but no post-baseline.

4.5.4.3. Sensitivity analyses

A sensitivity analysis will include death as an event rather than a censoring event for TTD. In this analysis patients will be counted as having event if they have a deterioration at two consecutive time points or die (before deterioration, or after the first deterioration but before the next assessment).

An additional analysis will use the time to first deterioration (TTFD) defined as the time from randomization to the date of the first increase in physical functioning score (≥ 5 points). The analyses specified in Section 4.4.1.3 will be repeated for these endpoints.

4.6. (Other) Safety Analyses

Not applicable.

4.7. Other Analyses

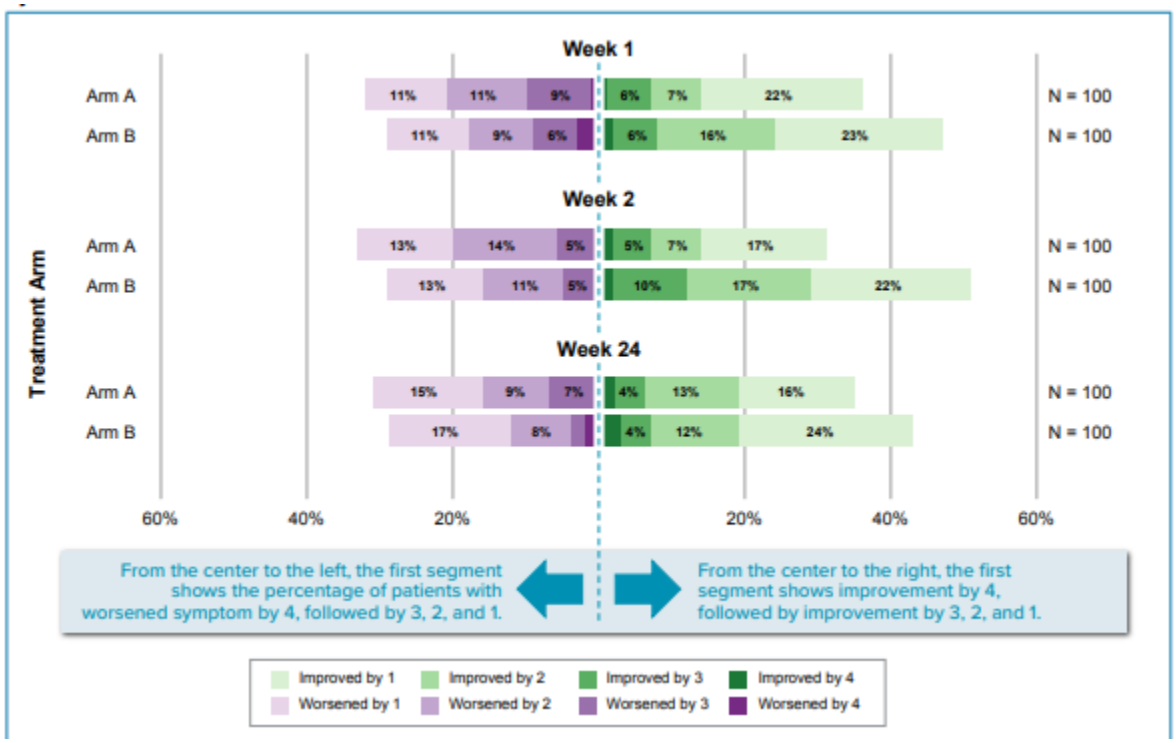
4.7.1. FACT-GP5 and PRO-CTCAE

4.7.1.1. Descriptive summaries

The numbers and percentages of patients reporting each category of frequency, severity, and interference (where reported) for each of the 13 PRO-CTCAE items will be summarized at each time point by treatment arm. The numbers and percentages reporting each category of FACT-GP5 and dichotomized as bothered/not bothered will also be summarized at each timepoint. The change from baseline in PRO-CTCAE items and FACT-GP5 per patient will be summarized at each time point by treatment arm.

The change from baseline in FACT-GP5, and all 13 PRO-CTCAE items will also be presented at each time point by treatment arm using the stacked bar charts (Figure 1) recommended by Basch et al (Basch et al., 2016, Zhou et al., 2018). All summaries and analyses will use the ITT and ITT-pembrolizumab populations.

Figure 1. Example stacked bar chart



The worst severity and interference with daily activities (highest score) achieved per patient during the study for decreased appetite, and fatigue will be summarized by treatment arm. This will also be reported using baseline subtraction to account for any residual level of patient-reported adverse events at baseline (i.e., the event will only be counted if the score is greater (worse) than that reported at baseline, if the score is the same or lower than baseline this is counted as zero). The worst severity and interference (without and with baseline subtraction) for decreased appetite and fatigue will be presented on a stacked bar chart by treatment arm.

The association between PRO-CTCAE decreased appetite and fatigue, and other PRO will be explored graphically. Plots of the mean NSCLC-SAQ total score and each symptom, and EORTC QLQ-C30 physical functioning over time will be presented categorized by each patient's worst (highest score) level of PRO-CTCAE decreased appetite and fatigue severity and interference during the study (both without and with baseline subtraction).

4.7.1.2. Analysis methods

To explore the association between the FACT-GP5 item ("I am bothered by side effects of treatment") and PRO-CTCAE decreased appetite and fatigue (severity and interference with usual activities), scatterplots will be used to plot each patient's score on the FACT-GP5 against the corresponding PRO-CTCAE item over all assessments, by treatment arm and overall, to enable a visual assessment of any association. The number and percentage of patients in each response category of severity of the PRO-CTCAE item (none, mild, moderate, severe, very severe) will be tabulated by the corresponding FACT-GP5 (bothered/not bothered) at that assessment, over all assessments. The association between the two items over all assessments

will be evaluated using an ordinal chi-square test. This will be repeated for the interference with usual or daily activities response of the PRO-CTCAE item (not at all, little bit, somewhat, quite a bit, very much).

The proportion of subjects who respond "bothered" to the GP5 item, dichotomized as "not bothered" (responses of 0 or 1) or "bothered" (response of 2, 3 or 4), will be analyzed with a binomial longitudinal (repeated measures) model using generalized estimating equations (GEEs), an extension of generalized linear models accounting for multiple observations in each subject, to estimate the odds of being bothered by the side effects of treatment (Stokes et al., 2012).

The analysis will be conducted using data from on-treatment visits. The model will only be fitted for on-treatment visit data where there is one (or more) patient(s) in both the "bothered" and "not bothered" categories for both treatment arms. A robust statistical fit for the GEE model requires sufficient data (in both the "bothered" and "not bothered" categories) due to constraints of statistical modelling process.

The dependent variable in the models will be the response of the GP5 item (proportion of patients bothered) at each study visit. The model will contain baseline GP5 score (six levels ranging from 0 [not at all] to 4 [very much] and missing), treatment arm, study visit (as categorical variables), and the randomization stratification factors as the independent variables. Participant will be a repeated measure (repeated across visits). An unstructured correlation matrix will be used.

The estimate of the treatment effect, once exponentiated, will be a ratio such that a value <1 indicates lower odds of being bothered by the side effects of treatment with selpercatinib compared to platinum-based and pemetrexed therapy with or without pembrolizumab.

If the initially specified model does not fit, the following strategies will be employed:

- Changing to an exchangeable correlation matrix
- Removing the stratification factors from the GEE model

Amending the latest visit included in the model to the visit with at least 10 patients in both treatment arms, where there are two (or more) patients in both the "bothered" and "not bothered" categories for both the arms.

The response of the GP5 will be alternatively dichotomized as follows:

- Bothered (i.e., the score is "1" ("a little bit") or higher).
- Not bothered (i.e., the score is "0" ("not at all")

The number of patients contributing to the overall model and at each timepoint will be presented together with the following statistics from the model:

Overall estimate of:

- Odds ratio and 95% CI
- P-value for the odds ratio between treatment arms; and
- By-visit estimates of:
    - Number and proportion of "bothered" patients for each treatment arm

4.7.2. Relationship between PROs and clinical endpoints

Analysis of relationships between PRO endpoints and clinical endpoints will be explored using graphical approaches.

The key relationships of interest to explore are:

Associations between best overall response (complete response-CR, partial response-PR, or stable) and the following PRO endpoints:

- time to confirmed deterioration in each NSCLC-SAQ symptom and total score
- time to confirmed deterioration of physical function
- time to confirmed deterioration of other HRQoL (EORTC QLQ-C30 subscales)

Kaplan-Meier plots of each of the outcomes above, stratified by best overall clinical response will be presented. A log-rank test will be used to compare the time to confirmed deterioration of each endpoint between the best overall response categories.

The association between baseline PRO scores and PFS (by BICR) and overall survival (OS; if OS data is available at the time of when PRO SAP is executed) will be evaluated using a Cox proportional hazards regression analysis including the relevant baseline PRO score as a covariate. This will use the analysis methods described in the J2G-MC-JZJC Clinical SAP for PFS (and OS if available) and will include each NSCLC-SAQ symptom score, total score, EORTC QLQ-C30/IL19 physical function and other EORTC QLQ-C30 subscales. Kaplan-Meier plots of PFS (and OS if available) for each baseline PRO score (dichotomized by the median value) will be presented.

4.7.3. Meaningful change analysis

The NSCLC-SAQ (McCarrier et al., 2016) will feature as a secondary endpoint to evaluate time to deterioration in symptoms. Time to deterioration (TTD) will be evaluated for each symptom using the relevant NSCLC-SAQ scale (i.e., Cough, Pain, Dyspnea, Fatigue, and Appetite). Therefore, meaningful change thresholds are required for each NSCLC-SAQ scale to establish TTD for secondary endpoint analyses. To date, meaningful within-person change thresholds have only been published for the NSCLC-SAQ Total score (Williams et al., 2022). Given this, psychometric analyses will be performed to establish the clinically meaningful change in symptom severity for patients in this study. This will include establishing meaningful score interpretation thresholds for the NSCLC-SAQ Total score and scale scores (i.e., Cough, Pain, Dyspnea, Fatigue, and Appetite). All psychometric analyses will be performed in the Psychometric Evaluation analysis set (in participants blinded to treatment arm allocation on a cut of the data prior to final lock of the database).

4.7.3.1. Interpretation of scores

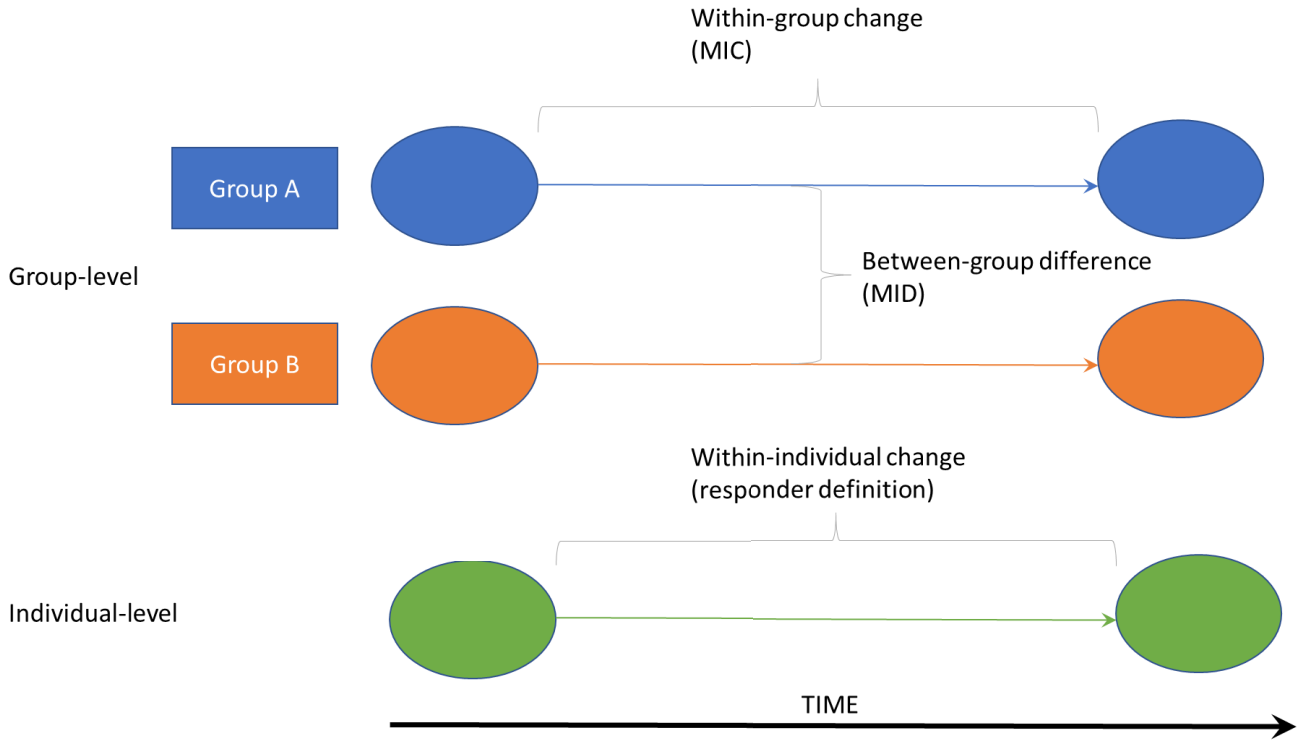
Score interpretability characterizes how we attribute meaning to observed changes and differences in scores, beyond what is provided for by “statistically significant” results.

4.7.3.2. Terminology

Terminology related to meaningful differences and changes in PRO scores varies considerably, and no consensus has been gained (King, 2011, Coon and Cappelleri, 2016). It is therefore important to define the types of difference or change associated with each term used throughout the remainder of this statistical analysis plan.

This study considers three ways in which scores on a PRO instrument can be interpreted (Figure 2). Two are at the group-level while the other is at the individual-level.

Figure 2. Terminology related to meaningful differences and changes in PRO scores



Abbreviations: MIC, Minimal Important Change; MID, Minimal Important Difference.

This figure provides a simple conceptual overview of meaningful change and related terminology. Coloured nodes represent PRO scores at hypothetical timepoints (e.g., leftmost blue node = Timepoint 1 [T1], rightmost blue node = Timepoint 2 [T2]); the difference between T1 and T2 at the Group-level defines the MIC; difference between MICs for Group A and Group B define the MID; and the difference between T1 and T2 at the Individual-level defines the Within-individual change.

4.7.3.3. Anchor-based methods

In anchor-based approaches, an easily interpretable and patient-centered external indicator is used to identify patients who have experienced an improvement on the concept being measured.

All anchor-based analyses will be performed in the psychometric analysis population using data from Day 1 of an on-treatment timepoint (for example, Day 1 Cycle 6 [Week 18] - this timepoint is suggested because it aligns with the meaningful change analyses performed in

Williams et al. (2022) and was therefore deemed to be beneficial in facilitating direct comparisons of the meaningful change thresholds for the NSCLC-SAQ derived in this study. Furthermore, according to the clinical trial SAP, 18 weeks is approximately half the median PFS, and therefore it is consistent with the general aim of anchor selection, where sufficient time has elapsed for change to have occurred, while sufficient participants remain in the study to make robust estimates of meaningful change).

4.7.3.3.1. Assessing potential anchors

The suitability of proposed anchors will be tested using polyserial (for NSCLC-SAQ Total score) and polychoric (for NSCLC-SAQ symptom scale scores) correlation coefficients to establish the relationship between the anchor categories and change in NSCLC-SAQ scores. Anchors with correlations of <0.3 will not be taken forward for analysis (Revicki et al., 2008).

4.7.3.3.2. Proposed anchor groups defining change

Each anchor that exhibits a sufficient strength of relationship with the PRO scores will be used to define groups of patients who have experienced improvement, no change, or worsening. Anchors that do not exhibit sufficient relationship strength may not be taken forward. Each potential anchor, as well as its categorical groupings, are defined in Table 8.

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
1	*Karnofsky Performance Scale Index*	*Total* *Fatigue* *Pain*	• *Major improvement: ≥30-point increase* • *Moderate improvement: 20-point increase* • *Minimal improvement: 10-point increase* • *No change: 0-point increase/decrease* • *Minimal worsening: 10-point decease* • *Moderate worsening: 20-point decrease* • *Major worsening: ≥ 30-point decrease*	Guided by Askew et al. (2009) where 10 points used as a meaningful threshold (Askew et al., 2009)
2	*ECOG*	*Total* *Fatigue* *Pain*	• *Moderate-to-Major improvement: 2-point decrease* • *Minimal improvement: 1-point decrease* • *No change: 0-point increase/decrease* • *Minimal worsening: 1-point increase*	

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			• *Moderate worsening: 2-point increase* • *Major worsening: ≥ 3-point increase*	
3	*Clinician-reported AEs – Cardiac Impairment*	*Dyspnea,*	• *Moderate-to-Major improvement: ≥ 2-grade decrease* • *Minimal improvement: 1-point decrease* • *No change: 0-grade increase/decrease* • *Minimal worsening: 1-grade increase* • *Moderate-to-Major worsening: ≥ 2-grade increase*	Guided by Musoro et al. (2020) (Musoro et al., 2020)
4	*Clinician-reported AEs – Respiratory disorders*	*Dyspnea*	• *Moderate-to-Major improvement: ≥ 2-grade decrease* • *Minimal improvement: 1-point decrease* • *No change: 0-grade increase/decrease* • *Minimal worsening: 1-grade increase*	Guided by Musoro et al. (2020) (Musoro et al., 2020)Guided by Musoro et al. (2020) (Musoro et al., 2020)

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			• *Moderate-to-Major worsening: ≥ 2-grade increase*	
5	*EORTC QLQ-C30 (Global Health Status [GHS])*	*Total, Fatigue, Dyspnea, Pain, Appetite Loss*	• *Major improvement: ≥ 24.9-point increase* • *Moderate improvement: ≥ 16.6 and <24.9-point increase* • *Minimal improvement: ≥ 8.3 and < than 16.6-point increase* • *No change: 0-point increase/decrease* • *Minimal worsening: ≥ 8.3 and < than 16.6-point decrease* • *Moderate worsening: Moderate improvement: ≥ 16.6 and <24.9-point decrease* • *Major worsening: ≥ 24.9-point decrease*	*Guided by possible raw score state changes (Cocks and Buchanan, In Press)**

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
6	*EORTC QLQ-C30 (Physical Functioning [IL19])*	*Total, Fatigue*	• *Major improvement: ≥ 20.1-point increase* • *Moderate improvement: ≥13.4 and < 20.1-point increase* • *Minimal improvement: ≥6.7 and < 13.4-point increase* • *No change: 0-point increase/decrease* • *Minimal worsening: ≥6.7 and < 13.4-point decrease* • *Moderate worsening: ≥13.4 and < 20.1-point decrease* • *Major worsening: ≥ 20.1-point decrease*	*Guided by possible raw score state changes (Cocks and Buchanan, In Press)**.
7	*EORTC QLQ-C30 (Fatigue)*	*Fatigue*	• *Major improvement: ≥ 33-point decrease* • *Moderate improvement: ≥ 22.2 and <33.3-point decrease*	*Guided by possible raw score state changes (Cocks and Buchanan, In Press)**

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			• *Minimal improvement: ≥ 11.1 and <22.2-point decrease* • *No change: 0-point increase/decrease* • *Minimal worsening: ≥ 11.1 and <22.2-point increase* • *Moderate worsening: ≥ 22.2 and <33.3-point increase* • *Major worsening: ≥ 33.3-point increase*	
8	*EORTC QLQ-C30 (Pain)*	*Pain*	• *Major improvement: ≥ 50.1-point decrease* • *Moderate improvement: ≥33.4 and < 50.1-point decrease* • *Minimal improvement: ≥16.7 and < 33.4-point decrease* • *No change: 0-point increase/decrease*	*Guided by possible raw score state changes (Cocks and Buchanan, In Press)**

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			• *Minimal worsening: ≥16.7 and < 33.4-point increase* • *Moderate worsening: ≥33.4 and < 50.1-point increase* • *Major worsening: ≥ 50.1-point increase*	
9	*EORTC QLQ-C30 (Dyspnea)*	*Cough*	• *Major improvement: ≥99.9-point decrease* • *Moderate improvement: ≥ 66.6 and <99.9-point decrease* • *Minimal improvement: ≥33.3 and < 66.6-point decrease* • *No change: 0-point increase/decrease* • *Minimal worsening: improvement: ≥33.3 and < 66.6-point increase* • *Moderate worsening: ≥ 66.6*	*Guided by possible raw score state changes (Cocks and Buchanan, In Press)**

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			and <99.9-point increase • *Major worsening: ≥99.9-increase*	
10	*EORTC QLQ-C30 (Appetite Loss)*	*Appetite*	• *Major improvement: ≥99.9-point decrease* • *Moderate improvement: ≥ 66.6 and <99.9-point decrease* • *Minimal improvement: ≥33.3 and < 66.6-point decrease* • *No change: 0-point increase/decrease* • *Minimal worsening: improvement: ≥33.3 and < 66.6-point increase* • *Moderate worsening: ≥ 66.6 and <99.9-point increase* • *Major worsening: ≥99.9-increase*	*Guided by possible raw score state changes (Cocks and Buchanan, In Press)**

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
11	*PRO-CTCAE (Appetite and Fatigue Severity & Interference items – 4 items total)*	*Appetite Loss, Fatigue*	• *Major improvement: ≤-3 point change* • *Moderate improvement: -2 point change* • *Minimal improvement: -1 point change* • *Stable: 0 point change* • *Minimal worsening: +1 point change* • *Moderate worsening: +2 point change* • *Major worsening: ≥3 point change*	
12	*GP5 Item*	*Total, Cough, Pain, Dyspnea, Fatigue, Appetite Loss*	• Major improvement: ≤-3-point change • Moderate improvement: -2-point change • Minimal improvement: -1-point change • Stable: 0 point change	

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			• Minimal worsening: +1-point change • Moderate worsening: +2-point change • Major worsening: ≥3-point change	
13	*EQ-5D-5L Items (Pain, Mobility, Usual Activities, Self-Care)*	*Total, Pain, Fatigue*	• Major improvement: ≥3 category improvement (e.g., extreme to slight) • Moderate improvement: 2 category improvement (e.g. extreme to moderate) • Minimal improvement: 1 category improvement (e.g., extreme to severe) • Stable: Same category (e.g., no to no) • Minimal worsening: 1 category	

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
			worsening (e.g., no to slight) • Moderate worsening: 2 category worsening (e.g., no to moderate) • Major worsening: ≥3 category worsening (e.g., no to severe)	

Table 8: Proposed anchors and groupings for defining meaningful change in NSCLC-SAQ

Anchor #	Anchor Measure	Relevant NSCLC-SAQ Scale(s)	Definition	Guiding literature (where available)
14	*EQ-5D-5L VAS*	*Total, Cough, Pain, Dyspnea, Fatigue, Appetite Loss*	• Major improvement: ≥ 21-point increase • Moderate improvement: ≥14 to <21-point increase • Minimal improvement: ≥7 to <14-point increase • Stable: change less than 7 points (negative and positive) • Minimal worsening: ≤-7 to >-14-point decrease • Moderate worsening: ≤-14 to >-21-point decrease • Major worsening: ≤-21-point decrease	Guided by Pickard et al. (2007) who estimated minimally important differences in cancer population (Pickard et al., 2007)

*The EORTC QLQ-C30 scoring scale is achieved through a linear transformation of its raw scores; therefore, only certain score change increments of its 0-100 scale are possible (see Cocks and Buchanan, In Press for further details (Cocks and Buchanan, In Press)).

**Cocks et al. (2012) (Cocks et al., 2012) and Koller et al. (2022) (Koller et al., 2022) provided group MID thresholds in NSCLC population, which were also consulted as a sense-check for appropriate meaningful values.

The frequency and percentage of patients within each group will be reviewed to assess suitability of the anchor. If very few patients (e.g., <10) are present in the minimal change groupings, anchor categories may be collapsed as additional outputs (e.g., combining minimal and moderate changes into a single category). Note that if collapsing minimal change anchor groups with higher magnitudes of change, the derived thresholds may no longer strictly reflect a *minimal* important change/difference, but instead reflect a more general threshold for importance (i.e., a clinically meaningful change).

4.7.3.3.3. Within-group mean change scores (MIC)

This analysis informs within-group MIC estimates. The mean change in PRO score will be calculated for patients classified according to [Anchors 1-13 in Table 8. The MIC estimate for each anchor will be derived using each groups mean change scores, focusing the estimates obtained from the minimal improvement and minimal worsening groups. In addition to the mean, the standard deviation, 95% CIs, median, minimum, maximum, 25th and 75th percentiles will be displayed. The mean change scores for each of the groups will be examined to ensure that they are monotonically increasing/decreasing, which is a further indication of an appropriate anchor. The frequency and percentage of patients within each group will also be reviewed to assess suitability of the anchor.

4.7.3.3.4. Between-group differences in mean change scores (MID)

This analysis informs between-group MID estimates. The mean change in PRO score will be calculated for patients classified as per **Error! Reference source not found.**.; the MID estimate for each anchor will be defined as the difference in mean change score between minimal worsening/minimal improvement and stable groups. Additionally, a linear regression model method will be used to estimate MIDs; change scores in the NLCLC-SAQ will be the dependent variable, with a binary anchor factor (i.e., stable = 0, and minimally improved/minimally worsened = 1), and the resulting slope parameter estimates for ‘improved’ and ‘deteriorated’ covariates being defined as the MIDs.

4.7.3.3.5. Within-individual change using Receiver Operating Characteristic (ROC) curves

This analysis informs a within-individual threshold (i.e., responder definition) estimates. ROC curve analysis will be used to find the change score that optimally discriminates between ‘worsened’ and ‘stable’ anchor categories, as well as ‘improved’ and ‘stable’ categories. Note that for this analysis, minimal, moderate, and major worsened groups will be collapsed into a single Worsened group, with an ‘Improved’ group created from collapsing minimal, moderate, and major improved groups. Potential responder definitions (i.e., all possible change scores) will be evaluated by finding an optimal cut-point using ROC curves where sensitivity and specificity are considered equally important, using indexes that aim to optimise both when neither are prioritised.

This will be summarized using ROC curves that plot sensitivity on the y-axis and one-minus-specificity on the x-axis for each 1-point change on the score. The area under the curve (AUC) will be reported for each curve, including 95% confidence intervals (CIs). If the AUC CIs contain 0.5, the score is no better than chance at predicting change as defined by the anchor. The proposed responder definition estimates will be evaluated by examining how close

corresponding change scores are to the top left corner of the curve which maximizes true positives while minimizing false positives.

To define the responder definition two different methods (Froud and Abel, 2014) will be considered to determine the threshold values:

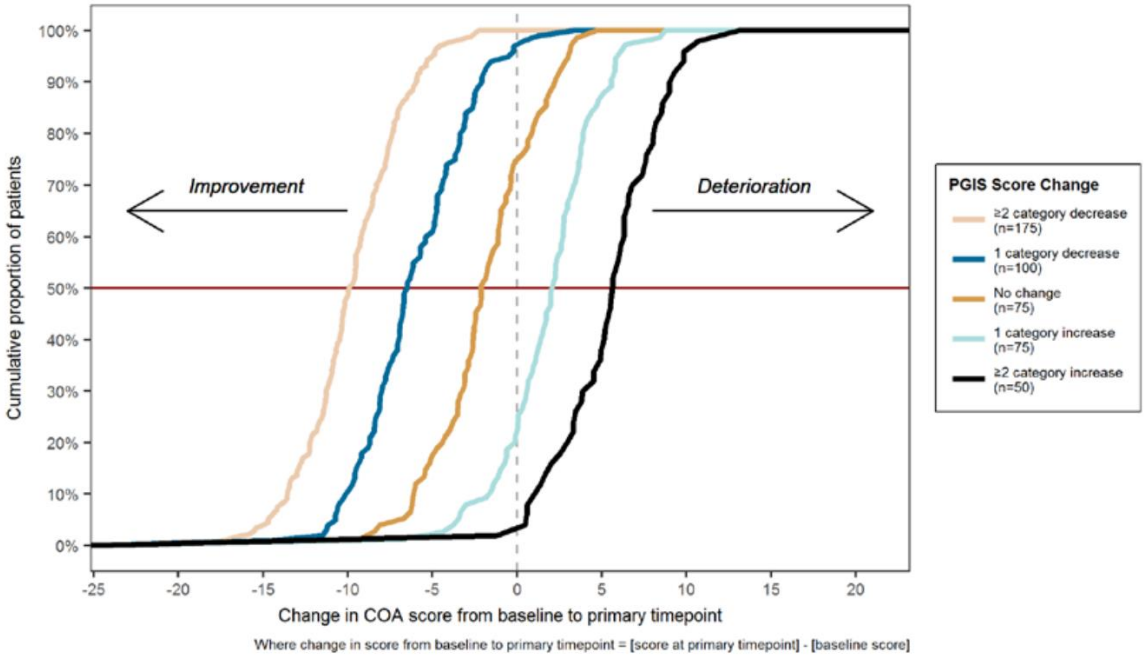
- Max[sensitivity+specificity-1], i.e. Youden's J index (Youden, 1950)
- Min[(1 - sensitivity)^2+(1 - specificity)^2], i.e., the sum of squares method.

Where the above methods identify alternative thresholds (i.e., two optimal thresholds for the same anchor), minimizing the sum of squares will be prioritized as this as this is proven mathematically to be closest to the top-left corner of the ROC curve (Froud and Abel, 2014).

4.7.3.3.6. Within-individual change using cumulative distribution function (CDF) and probability density function (PDF) curves

As the analyses proposed above may suggest different responder definitions, the CDF will allow all proposed responder definitions for an anchor to be evaluated simultaneously. Notably, the aim of CDF plots is not to estimate responder definitions but to evaluate the performance of each; CDF plots allow specific the consequences of setting specific within-patient change thresholds through assessment of the proportion of improved/worsened individuals that are correctly indicated. This is consistent with current regulatory guidance that suggests exploring consequences of establishing meaningful change thresholds (Food and Drug Administration, 2018).

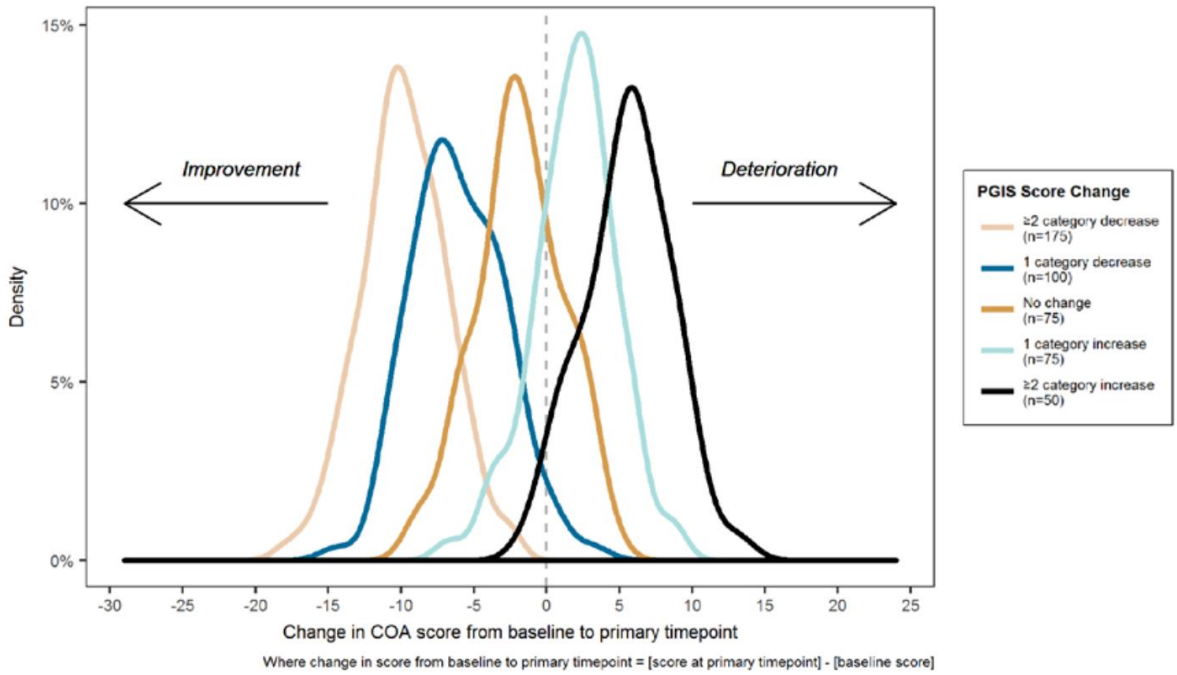
Figure 3: Example of Empirical Cumulative Distribution Function (eCDF) Curves of Change in COA Score from Baseline to Primary Time Point by Change in Patient Global Impression of Severity (PGIS) score (Food and Drug Administration, 2018)



The CDF plots will display a continuous PRO score change from baseline on the x-axis for each of the scores. The axis will be ordered from the best possible improvement on the left through to the worst possible deterioration on the right. This ordering will be the same for all scores, considering the directionality of scores in relation to health. The cumulative percentage of patients experiencing that change will be displayed on the y-axis, plotted as an empirical (i.e., stepwise) CDF.

The CDF curves will be split by anchor group (according to each anchor specified in **Table 8**) so the separation of the curves can be visually compared across the range of potential thresholds identified across the different anchor methods. Additionally, the points on each curve corresponding to a cumulative 50% of responders will be highlighted.

Figure 4: Example of Density Function (PDF; often estimated using kernel density estimation) Curves of Change in COA Score from Baseline to Primary Time Point by Change in PGIS Score (Food and Drug Administration, 2018)



Probability density functions (PDFs) will also be plotted with continuous PRO score change from baseline on the x-axis for each of the scores, from the best possible improvement on the left through to the worst possible deterioration on the right. The percentage of patients experiencing that change will be displayed on the y-axis, plotted as a kernel-smoothed PDF. The PDF curves will be split by anchor group (according to each anchor specified in **Table 8**) so the separation of the curves can be visually compared across the range of potential thresholds identified across the different anchor methods.

4.7.3.4.Distribution-based methods

A distribution-based approach will be employed, where distributional properties of the scores themselves are used to guide potential responder definitions estimated from anchor-based

approaches (FDA, 2009, McLeod et al., 2011), identifying the amount of change that exceeds measurement error. The distribution-based methods will consist of computing the following:

- 0.5 of a standard deviation (SD): Research has suggested that the 0.5 SD estimate may approximate a responder definition for PRO scores (Norman et al., 2003). Moreover, empirical evidence tends to converge on the 0.5 SD criteria as being meaningful to patients (Osoba, 2004). This distribution based approach will involve calculating 0.5 of the SD at Baseline. To provide additional sensitivity analysis information, values of 0.2 and 0.8 of the SD at Baseline will also be calculated.
- SEM: The SEM is calculated as the standard deviation at Baseline multiplied by the square root of one minus the reliability of the score at baseline [SD * $(1\text{-}r)^{1/2}$]. Therefore, the SEM is equivalent to 0.5 SD when reliability equals 0.75, and decreases as reliability increases. The internal consistency, Cronbach's alpha, at baseline (cycle 1, day 1) will be used for the reliability for NSCLC-SAQ Total score (Wyrwich et al., 1999). For NSCLC-SAQ scores with 2 items (i.e., Pain and Fatigue), the Spearman-Brown formula will be used as a more appropriate measure of reliability (Eisinga et al., 2013). For NSCLC-SAQ scores with 1 item (i.e., Cough, Dyspnea, and Appetite), the ICC will be used from a test-retest analysis of stable participants. Stability will be defined having not experienced meaningful change in EORTC QLQ-C30 (GHS) and EQ-5D-5L VAS (according to thresholds defined in **Error! Reference source not found.**) between Cycle 2 and Cycle 3. A value of 1 SEM will be used as the estimate of the responder threshold.

4.7.3.5.Triangulation of estimates

The application of multiple methods to estimate an MID or responder definition for a PRO score in a specific patient population will almost always result in a range of values for the MID (Revicki et al., 2006). Triangulation is the process of bringing all the estimates together with the aim of converging on a single value or small range of values.

There is no consensus on a 'best' process to conduct triangulation; however, using recommendations from the limited literature available (Revicki et al., 2006, Revicki et al., 2008, Coon and Cook, 2018), the following process will be employed:

- The different MID (or responder definition) estimates with 95% CIs as available are visualized on a graph to depict the range of estimates to identify an appropriate range of values.
- When considering the hierarchy of estimates from the different approaches, anchor-based estimates should be assigned more weight than distribution-based estimates, as they are more closely related to the patient perspective.
- Specific to anchor-based analyses, the proximity of anchor to the target PRO instrument should also be considered, with more importance to estimates generated from more closely linked concepts and highly correlated pairings.

Additionally, to guide triangulation of estimates (but not explicitly define them), a correlation-weighted average (Harper et al., 2018, Trigg and Griffiths, 2021) of estimates may be calculated,

where estimates are weighted by the observed correlations between change in anchor and score as follows:

$$M_{weighted} = \frac{\sum_{i=1}^{n} |r_i||x_i|}{\sum_{i=1}^{n} |r_i|}$$

where x denotes each [absolute] estimate and r denotes the [absolute] correlation coefficient of each anchor-scale combination, for each i of n total estimates.

4.7.3.6. Alternative meaningful change threshold establishment

Meaningful change thresholds are planned to be determined for NSCLC-SAQ domain scores and Total score, with the Total score threshold being compared to those derived in Williams et al. (2022) (Williams et al., 2022). However, idiosyncrasies in individual trial data do not always permit robust meaningful change threshold estimation (e.g., due to low correlations between NSCLC-SAQ and potential anchors). If the available trial data does not permit robust estimation of anchor-based meaningful change thresholds, as described in Section 4.7.3.3, alternative methods and sources of information may be employed to define meaningful change thresholds for the purposes of secondary endpoint analysis in LIBRETTO-431. Sources of alternative meaningful change threshold establishment may include sole consideration of distribution-based methods (see Section 4.7.3.4), use of 10% of instrument range (Ringash et al., 2007) use of 15% of scale range (IQWiG, 2020), or thresholds derived in Williams et al. (2022) (Williams et al., 2022).

4.8. Subgroup Analyses

Not applicable.

4.9. Interim Analyses

4.9.1. Data Monitoring Committee (DMC) or Other Review Board

Not applicable.

4.10. Changes to Protocol-Planned Analyses

Not applicable.

5. Sample Size Determination

Not applicable.

6. Supporting Documentation

6.1. Appendix 1: Demographic and Baseline Characteristics

Not applicable.

6.2. Appendix 2: Treatment Compliance

Not applicable.

6.3. Appendix 3: Clinical Trial Registry Analyses

Not applicable.

7. References

Askew, R. L., Y. Xing, J. L. Palmer, D. Cella, L. A. Moye and J. N. Cormier (2009). "Evaluating minimal important differences for the FACT-Melanoma quality of life questionnaire." Value in Health **12**(8): 1144-1150.

Basch, E., L. J. Rogak and A. C. Dueck (2016). "Methods for implementing and reporting patient-reported outcome (PRO) measures of symptomatic adverse events in cancer clinical trials." Clinical therapeutics **38**(4): 821-830.

Bell, M. L. and D. L. Fairclough (2014). "Practical and statistical issues in missing data for longitudinal patient-reported outcomes." Stat Methods Med Res **23**(5): 440-459.

Cocks, K. and J. Buchanan (In Press). "How scoring limits the usability of minimal important differences (MIDs) as responder definition (RD) - An exemplary demonstration using EORTC QLQ-C30 subscales." Quality of Life Research.

Cocks, K., M. King, G. Velikova, G. de Castro Jr, M. M. St-James, P. Fayers and J. Brown (2012). "Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30." European journal of cancer **48**(11): 1713-1721.

Cocks, K., M. T. King, G. Velikova, M. Martyn St-James, P. M. Fayers and J. M. Brown (2011). "Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30." Journal of Clinical Oncology **29**(1): 89-96.

Coffman, C. J., D. Edelman and R. F. Woolson (2016). "To condition or not condition? Analysing 'change' in longitudinal randomised controlled trials." BMJ Open **6**(12): e013096.

Coon, C. D. and J. C. Cappelleri (2016). "Interpreting Change in Scores on Patient-Reported Outcome Instruments." Therapeutic Innovation & Regulatory Science **50**(1): 22-29.

Coon, C. D. and K. F. Cook (2018). "Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores." Qual Life Res **27**(1): 33-40.

Eisinga, R., M. Grotenhuis and B. Pelzer (2013). "The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?" Int J Public Health **58**(4): 637-642.

Fairclough, D. L. (2002). Design and analysis of quality of life studies in clinical trials, Chapman and Hall/CRC.

Fairclough, D. L. (2010). Design and analysis of quality of life studies in clinical trials, Chapman and Hall/CRC.

Fayers, P. M., N. K. Aaronson, K. Bjordal, M. Groenvold, D. Curran and A. Bottomley (2001). The EORTC QLQ-C30 Scoring Manual, Brussels: European Organisation for Research and Treatment of Cancer.

FDA (2009) "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims." Guidance for Industry.

Food and Drug Administration (2018). "Patient-focused drug development guideance public workshop. Methods to identify what is important to patients & select, develop or modify fit-for-purpose clinical outcomes assessments (https://www.fda.gov/media/116277/download)."

Froud, R. and G. Abel (2014). "Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status." PLOS ONE **9**(12): e114468.

Harper, A., C. Trennery, K. Sully and A. Trigg (2018). Triangulating estimates of meaningful change or difference in patient-reported outcomes: application of a correlation-based weighting procedure. Quality of Life Research.

IQWiG (2020). General Methods Version 6.0, Institut fur Qualitat under Wirtschaftlichkeit im Gesundheitswesn.

King, M. T. (2011). "A point of minimal important difference (MID): a critique of terminology and methods." Expert Rev Pharmacoecon Outcomes Res **11**(2): 171-184.

Koller, M., J. Z. Musoro, K. Tomaszewski, C. Coens, M. T. King, M. A. Sprangers, M. Groenvold, K. Cocks, G. Velikova and H.-H. Flechtner (2022). "Minimally important differences of EORTC QLQ-C30 scales in patients with lung cancer or malignant pleural mesothelioma–Interpretation guidance derived from two randomized EORTC trials." Lung Cancer **167**: 65-72.

McCarrier, K. P., T. M. Atkinson, K. P. DeBusk, A. M. Liepa, M. Scanlon, S. J. Coons, P.-R. O. Consortium and N. S. C. L. C. W. Group (2016). "Qualitative development and content validity of the Non–Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ), a patient-reported outcome instrument." Clinical therapeutics **38**(4): 794-810.

McLeod, L. D., C. D. Coon, S. A. Martin, S. E. Fehnel and R. D. Hays (2011). "Interpreting patient-reported outcome results: US FDA guidance and emerging methods." Expert Rev Pharmacoecon Outcomes Res **11**(2): 163-169.

Musoro, J. Z., C. Coens, S. Singer, S. Tribius, S. F. Oosting, M. Groenvold, C. Simon, J. P. Machiels, V. Grégoire and G. Velikova (2020). "Minimally important differences for interpreting European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 scores in patients with head and neck cancer." Head & neck **42**(11): 3141-3152.

Norman, G. R., J. A. Sloan and K. W. Wyrwich (2003). "Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation." Med Care **41**(5): 582-592.

Osoba, D. (2004). The clinical value and meaning of health-related quality-of-life outcomes in oncology. Outcomes Assessment in Cancer: Measures, Methods and Applications. C. C. Gotay, C. Snyder and J. Lipscomb. Cambridge, Cambridge University Press**:** 386-405.

Pickard, A. S., M. P. Neary and D. Cella (2007). "Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer." Health and quality of life outcomes **5**(1): 1-8.

Revicki, D., R. D. Hays, D. Cella and J. Sloan (2008). "Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes." J Clin Epidemiol **61**(2): 102-109.

Revicki, D. A., D. Cella, R. D. Hays, J. A. Sloan, W. R. Lenderking and N. K. Aaronson (2006). "Responsiveness and minimal important differences for patient reported outcomes." Health and Quality of Life Outcomes **4**: 70-70.

Ringash, J., B. O'Sullivan, A. Bezjak and D. A. Redelmeier (2007). "Interpreting clinically significant changes in patient-reported outcomes." Cancer **110**(1): 196-202.

Stokes, M. E., C. S. Davis and G. G. Koch (2012). Categorical Data Analysis Using SAS, Third Edition. Cary, NC, SAS Institute Inc.

Trigg, A. and P. Griffiths (2021). "Triangulation of multiple meaningful change thresholds for patient-reported outcome scores." Quality of Life Research **30**(10): 2755-2764.

Williams, P., T. Burke, J. M. Norquist, C. Daskalopoulou, R. M. Speck, A. Samkari, S. Eremenco and S. J. Coons (2022). "Non-Small Cell Lung Cancer Symptom Assessment Questionnaire (NSCLC-SAQ): Measurement Properties and Estimated Clinically Meaningful Thresholds From a Phase 3 Study." JTO clinical and research reports **3**(4): 100298.

Wyrwich, K. W., W. M. Tierney and F. D. Wolinsky (1999). "Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life." J Clin Epidemiol **52**(9): 861-873.

Youden, W. J. (1950). "Index for rating diagnostic tests." Cancer **3**(1): 32-35.

Zhou, X., D. Eid and A. Gnanasakthy (2018). "Methods for reporting the patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE) data in cancer clinical trials." Value in Health **21**: S226.

Signature Page for VV-CLIN-070444 v2.0

Approval	PPD 27-Apr-2023 18:30:38 GMT+0000

Signature Page for VV-CLIN-070444 v2.0