Statistical Analysis Plan (SAP) for early toxicity reporting for the NARLAL II trial

This SAP has been created using a layout recommended by "Liverpool clinical Trial centre - LCTC":

https://www.lctc.org.uk/Content/SAP%20Statement_%20Elaboration%20Document%20v1.0.pdf

The form is also linked to from the Equator network:

https://www.equator-network.org/reporting-guidelines/guidelines-for-the-content-of-statistical-analysis-plans-in-clinical-trials/

Section 1: Administrative information

Part 1: Title and Trial registration

Statistical analysis plan for the NARLAL II trial (Novel Approach to Radiotherapy in Locally Advanced Lung cancer - Heterogeneous FDG-guided dose escalation with concomitant Navelbine), clinicaltrials.gov NCT02354274. This specific SAP describes the analysis of early toxicity (within 6 months after patient randomisation) related to treatment in the NARLAL II trial. The SAP is based on the NARLAL II primary endpoint SAP (dated 13th February 2023).

Part 2: SAP Version

The current document is the first version of the final statistical analysis plan for reporting of early toxicity outcomes, finalised after the end of patient accrual but prior to data unblinding. The date of the current version (version 1.0) is the 27th November 2023.

Part 3: Protocol Version

The basis for this SAP is the English translation of the Danish protocol (version 6) as approved by the Danish Research Ethics Committee (REC) on the 15th August 2022.

Part 4: SAP Revisions – revision history, with justification and timing

The current SAP is the first version of the SAP; thus, there are no revisions.

Part 5 Roles and Responsibility – non-signatory names and contribution

Names are listed alphabetically by first name:

Ane Appelt, Member of the working group drafting the SAP

Carsten Brink, Member of the working group drafting the SAP, and author of the R-package for data analysis

Charlotte Kristiansen, Member of the working group drafting the SAP

Lone Hoffmann, Member of the working group drafting the SAP

Marianne Marquard Knap, Member of the working group drafting the SAP

Mikkel Drøgemüller, Member of the early toxicity group

Rune Slot Thing, Member of the early toxicity group

Tarje Halvorsen, Member of the early toxicity group

Torben Schjødt Hansen, Member of the early toxicity group

Vilde Drageset Haakensen, Member of the early toxicity group

Part 6: Roles and Responsibility – signatures

Tine Schytte, Primary investigator

27th November 2023

Signature:

Section 2: Introduction

Part 7: Background and rationale

The background and rationale for the NARLAL II study are described in section one (*Background*) of the clinical protocol. This SAP specifically supports the analysis & reporting of early toxicity. The potential risk of increased radiation-induced toxicity is a major concern for radiotherapy dose escalation trials; hence, detailed toxicity reporting is essential.

Part 8: Objectives

The overall objective of this trial is to examine the effect of inhomogeneous, FDG-PET-driven escalation of radiation dose to the primary tumour and involved lymph nodes, compared to standard uniform dose, in definitive chemo-radiation treatment of inoperable locally advanced NSCLC (stage IIB-IIIB).

Secondary study objectives include evaluation of acute and late toxicity, overall and recurrence-free survival, and correlation of radiation dose with tumour and nodal control. Toxicity is scored according to CTCAE version 4.0. Acute and late toxicity will be separately evaluated, with the current SAP focusing on early toxicity.

Section 3: Trial Methods

Part 9: Trial design – description of trial design

The trial is a phase III multi-centre trial, randomising patients with locally advanced NSCLC between standard and inhomogeneous radiotherapy dose escalation. The randomisation ratio is 1:1. Toxicity was scored at each follow-up visit: scheduled at baseline, weekly during radiotherapy (week 2, 3, 4, 5, 6, and 7 after start of radiotherapy), every third month after randomisation for the first two years, every sixth month until five years, and once a year until ten years after randomisation.

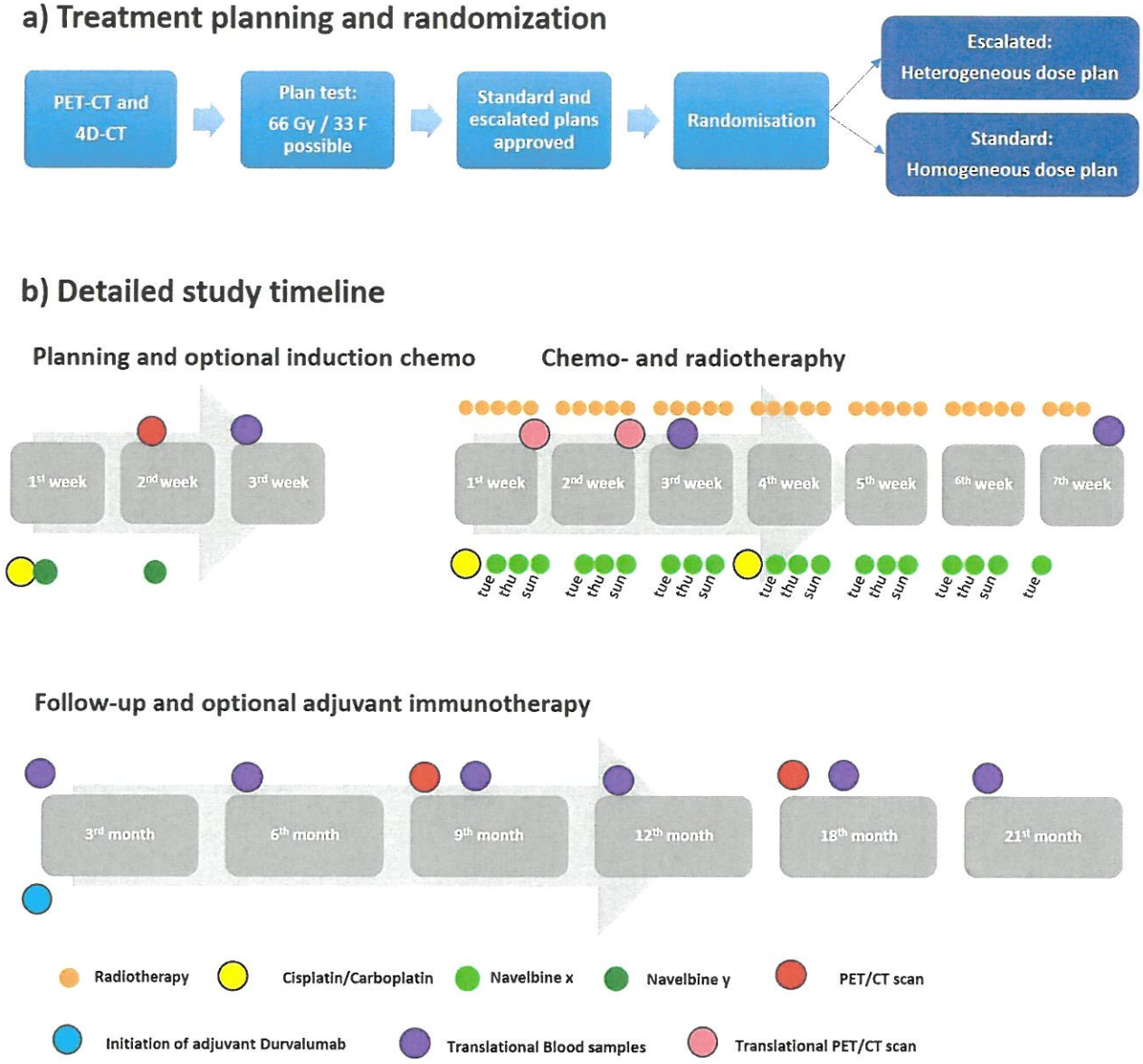


Figure 1: Study schedule for patients in the NARLAL2 trial.

a) Overall study schedule leading up to randomization.

b) Detailed time schedule of chemo- and radiation therapy for a patient who starts on Monday and the subsequent follow-up. Standard follow-up CTs are not shown. In 2019, adjuvant treatment with durvalumab was approved as part of the standard treatment of advanced lung cancer. The implementation in the NARLAL2 study schedule is indicated (blue). Translational PET/CT scans during RT as well as translational blood samples before, during and after RT are indicated.

Part 10: Randomisation

The randomisation was performed within blocks and was stratified on treatment institution and histology (squamous vs non-squamous cell carcinoma). The randomisation lists were created by the "Open Patient data Explorative Network" (OPEN – Region of Southern Denmark) and implemented in the related Redcap

database. The size of the stratification blocks was unknown to the investigators. The study was open-label for the patient and the treating physician, but the allocation to each treatment arm was blinded. For each centre, the randomisation was performed within the Redcap database, typically accessed by the local clinical research unit. The ability to deliver a standard arm treatment plan was a prerequisite for study enrolment and randomisation. Radiotherapy treatment plans for both study arms were finalised by the local dose planner and approved by the treating physician before randomisation was communicated to the treating team.

Part 11: Sample size

The trial sample size was based on the primary endpoint (locoregional control). At the time of protocol initiation, the number of patients planned for the study was 330. However, the introduction of Durvalumab (as consolidation treatment post chemo-radiotherapy for responders) during the study period resulted in a protocol amendment, including an update of the sample size calculation. Following the amendment, the study's final number of planned patients was 350 (175 per study arm).

The trial will continue recruiting after the inclusion of patient number 350 and until the study group is ready to publish the primary endpoint. This is to increase the power for detecting potential differences in (particularly late) toxicity; and has been implemented as a REC approved major amendment. Toxicity will be evaluated and reported for all patients enrolled on the trial. However, this SAP focuses on the reporting of toxicity occurring within 6 months after randomisation for the initial 350 patients alone.

Part 12: Framework

The trial was designed to have an identical lung toxicity profile in the two study arms. This was obtained by performing dual dose planning. Radiotherapy treatment plans for both study arms were finalised by the local dose planner and approved by the treating physician before randomisation was communicated to the treating team. For the standard, homogeneous dose plan, constraints were set for lungs, heart, oesophagus, spinal cord/spinal canal, and overall maximum dose. For the escalated, heterogeneous plan, additional constraints were set for brachial plexus, trachea, bronchi, aorta, all tissue in the mediastinum not included in other delineated organs at risk, and thoracic wall. All constraints were mandatory. Additionally, it was required that the mean lung dose in the two plans deviated less than ±1 Gy and that the volume of the lungs receiving 20 Gy differed less than ±2 %-point.

Part 13: Statistical Interim analyses and stopping guidance

Interim analyses have been carried out during the inclusion period following the specifications in the trial protocol. The main purpose of the interim analyses was to detect and pause the trial in case of unexpected and unacceptable acute or semi-acute/late toxicity or reduced overall survival in the dose escalation arm.

The interim toxicity analyses used the O'Brien-Fleming [1] method to define the stopping rules for acute and semi-acute/late toxicity, respectively, with an overall significance level of 5% for each set of analyses. The acute toxicity interims were performed after 3 months of follow-up of the initial 20, 40, and 60 patients included in the dose-escalation arm. The semi-acute/late analyses were performed for the same number of patients after one year of follow-up. The aim was to stop the study if radiation-related toxicity of grade 4+ was observed for more than 10% of the patients at three months and more than 5% at one year in the dose escalation arm.

Further details of the interim analyses are provided in the protocol. Due to legal requirements, the principal investigator must evaluate all serious adverse events (SAE) during the study. Thus, the principal investigator

had access to all SAE information and performed the toxicity-related interim analyses. All interim analyses were passed without pausing the study.

Part 14 Timing of final analysis

The analysis of early toxicity is planned for 6 months after inclusion of patient number 350. This analysis depends on complete information on the early toxicity endpoint at the trial management centre, and relevant additional information on patient characteristics and treatment. Full information includes:

- Key baseline patient characteristics;
- Information on treatment delivered, including any protocol violations;
- Radiotherapy plans transferred to the national treatment plan bank (DcmCollab) and ready for evaluation;
- Complete information on local recurrence within 6 months after randomisation
- Information on the administration of Durvalumab as consolidation treatment within 6 months after randomisation (as the lung toxicity may depend on Durvalumab);
- Survival status;
- Toxicity information for all treatment-related toxicities within 6 months after randomisation.

The trial database will indicate the date for the latest update of survival, toxicity, and prescription of Durvalumab for each patient. This date information will be used to ensure that all the requested information is available before analysis.

A database export for the analysis for the current SAP will be performed at the date when all the above data are available in the trial management centre and all centres report that they have finalised their local data curation.

Part 15: Timing of outcome assessments

All visits during radiotherapy and follow-up times are defined in the protocol (sections 8.3-8.4). Visits are scheduled at baseline, weekly during radiotherapy (week 2, 3, 4, 5, 6, and 7 after the start of radiotherapy), every third month after randomisation for the first two years, every sixth month until five years, and once a year until ten years after randomisation. Toxicity is scored at all visits. Lung function tests, and recurrence evaluation by CT are performed at follow-up visits. Serious adverse events (SAE) were collected separately as part of the study safety monitoring. All SAEs were evaluated by the local investigators. Retrospectively, all treatment-related SAEs have been evaluated to ensure that the treatment-related SAEs also have been scored as toxicity within the relevant follow-up visit. Handling of SAE-reported toxicity is discussed in more detail in Parts 26 and 27.

Section 4: Statistical Principles

Part 16 Confidence intervals and p-values

All applicable statistical tests will be 2-sided and performed using a 5% significance level; reported confidence intervals will be 95%.

Part 17: Description of any planned adjustment for multiplicity, and if so, including how the type 1 error is to be controlled

Potential differences in toxicity between the two treatment arms will be analysed for each type of toxicity, as well as for aggregated toxicity per organ system (lung, oesophageal, and gastrointestinal), during as well as after radiotherapy (up to 6 months). Consequently, a large number of tests for differences in toxicity

rates will be performed. All p-values will be presented without adjustment for multiplicity, as this provides a conservative evaluation of any safety signals. However, to guide the appraisal of the findings, all p-values will also be subjected to the Benjamini-Hochberg false discovery rate procedure (using a ratio between false positives and all positive results (Q) of 0.10) [2], and resulting significant findings will be indicated (see footnote for a description of the Benjamini-Hochberg procedure[A]). Adjustment of multiplicity is a field of much debate, and it can be argued that other corrections should have been used. However, this choice of correction was selected instead of e.g. the Bonferroni correction, which tends to reject too many true significant tests. But at the end of the day, the clinical impact of any potential differences will be important in evaluating the entire trial; thus, it has been decided not to perform even more advanced adjustments for multiplicity.

Part 18: Confidence intervals (CI) to be reported

Any confidence intervals on proportions (fractions of patients experiencing at least a given level of toxicity), including on figures, will be based on the assumption of a binomial distribution. The confidence intervals will be 95% confidence intervals.

Part 19. Adherence and Protocol Deviations

Radiotherapy-related protocol deviations will be presented in the publication if relevant for the analysis of (acute) toxicity.

Major protocol deviations are defined as the following:

- Patients partly or fully treated according to another arm than the one they were allocated to by randomisation
- Patients for which the randomisation result was known before the finalisation of both treatment plans (escalated and standard arm)
- Patients treated with a plan that did not obey the high-priority OAR constraints (spinal cord, bronchi, oesophagus, lung, and heart)

Lack of protocol treatment adherence is defined as:

- Patients who did not finish the intended radiotherapy treatment

A study flow diagram based on the recommendation from the CONSORT group will be provided and will include the number of deviations defined above (see Part 21, section 5).

Part 20: Analysis populations

The analysis of early toxicity will consider eligible study patients who have received at least one fraction of radiotherapy (i.e. who have initiated study treatment). This excludes patients who initially consented to the trial but were found to be ineligible post-consent, for whom radiotherapy planning according to trial guidelines was not possible, or who withdrew consent prior to treatment initiation. Data for patients who were withdrawn from the trial at a later timepoint (either due to withdrawn consent or a clinical decision not to complete curative radiotherapy) will only be included until the withdrawal date.

For this safety analysis, a 'per protocol' approach will be used; i.e. patients will be analysed based on the delivered treatment (and not the intention-to-treat arm). Consequently, patients treated in the opposite

---

A

https://stats.libretexts.org/Bookshelves/Applied_Statistics/Biological_Statistics_(McDonald)/06%3A_Multiple_Tests/6.01%3A_Multiple_Comparisons

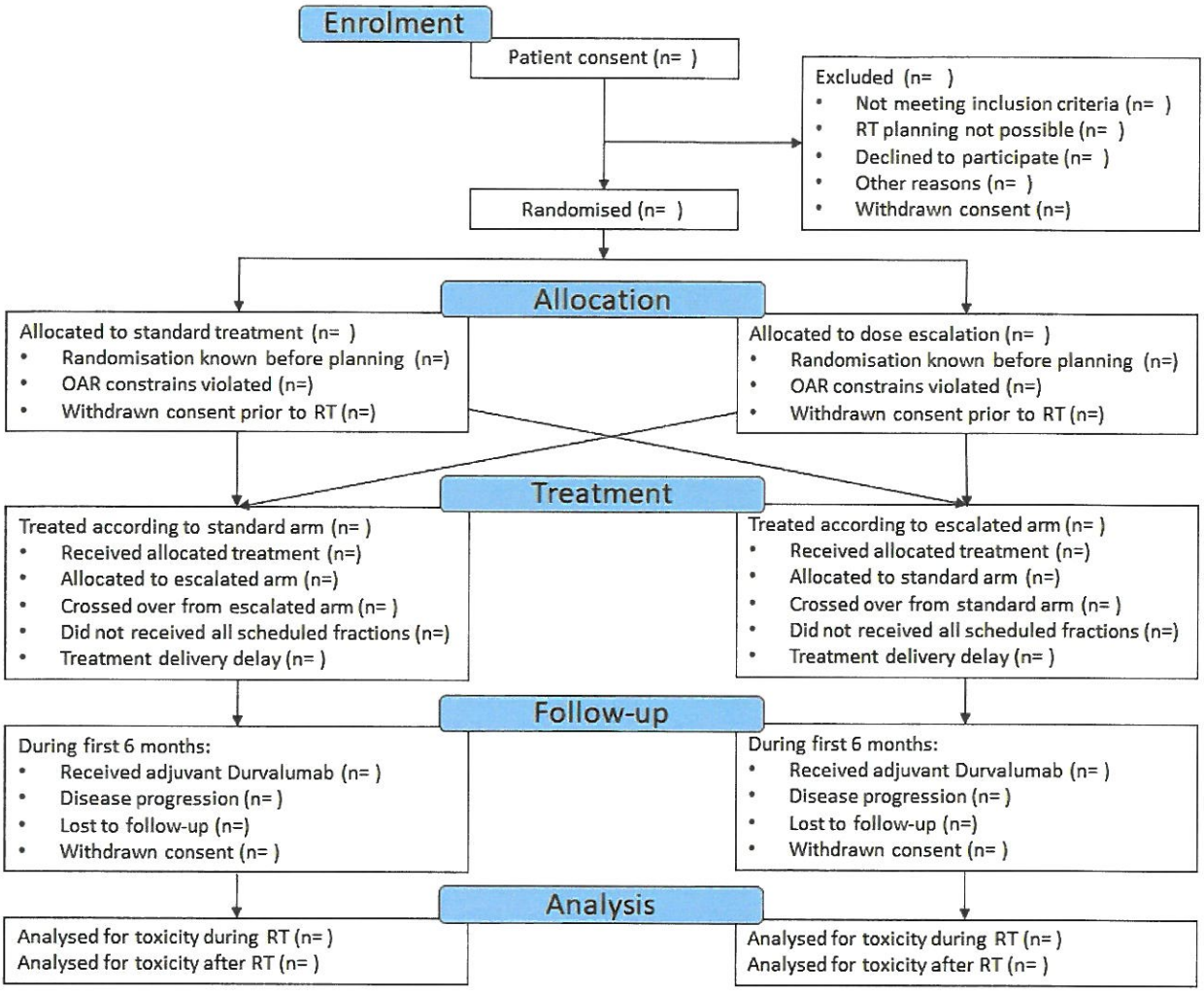
arm than indicated by the randomisation will be analysed in the arm of the actual treatment. Patients experiencing cross-over, i.e. change of treatment arm during the treatment course, will be analysed within the arm for which most of the treatment fractions were delivered. The reason for early treatment completion will be provided for patients not receiving all radiotherapy treatment fractions. Four categories will be used: chemo-radiotherapy related, patient decision, unknown, and other.

Section 5: Trial Population

Part 21:

All patients treated in the participating centres and fulfilling the inclusion criteria were candidates for the trial. The trial remained open during the COVID-19 pandemic, during which some centres had to reduce the trial capacity. Thus, not all potential candidates may have been offered trial participation. Systematic screening logs were not maintained at all centres; therefore, the number of patients screened for the trial will not be reported.

Figure 2: Flow diagram of patients included, treated and analysed on the trial. The layout is based on the recommendations from the CONSORT group http://www.consort-statement.org/consort-statement/flow-diagram, with modifications to fit the current trial. Major protocol deviations (defined in Part 19) should be included in the "allocation" section, number of consent withdrawals during follow-up should be mentioned in the "follow-up" section. For all relevant entries, the reason for a given deviation should, if possible, be documented in the caption of the figure or in the appendix. Treatment delivery delay is defined as treatment times beyond 53 days.



Part 24: Withdrawal/Follow-up – level of withdrawal

The number of patients that withdrew their consent will be reported separately for the two arms, at each relevant time point.

Part 25: Baseline patient & treatment characteristics

All patient and key treatment characteristics will be reported separately for the two treatment arms. An example layout is shown in Table 1 below, based on simulated data. Categorical and ordinal variables will be presented as numbers and percentages. Continuous values will be presented by their medians and interquartile ranges. Furthermore, cumulative and differential plots of all continuous variables will be plotted separately for the two treatment arms. The following variable will, as a minimum, be included in the patient & treatment characteristics:

1) Sex

2) Age
3) Histology
4) Stage
5) Performance status
6) Lung function (FEV1 and DLCO)
7) Smoking at the start of RT
8) Volume of GTV-T
9) Volume of GTV-N
10) Volume of PTV
11) No. of Navalbine (median [IQR]); i.e. number of days Navalbine was administered during RT
12) No. of series of platinum-based chemotherapy during RT (as categorical variable)
13) Mean dose to the lung, heart and oesophagus
14) Maximum dose (D_{1cm3}) to the heart, oesophagus, trachea and bronchi
15) Maximum dose ($D_{0.05cm3}$) to the spinal cord
16) Number of fractions delivered (<30, 30-32, 33)
17) Number of patients with treatment time beyond 53 days
18) Duvalumab (yes/no)

Since this is a randomised trial, no test is performed to evaluate whether the baseline variables of the two treatment arms are sampled from the same population.

The database includes the absolute value of FEV1 spirometry measures. For publication, the percentage of expected FEV1 will be reported; the reference values will be calculated using the equation in Løkke et al. [3]

Table 1: Example of the patient characteristics table based on simulated data and summarised using the statistical software developed as part of the SAP development. The simulated data are not sampled from a previous cohort; thus, the data shown might be very different from an actual cohort. In particular, the simulated randomisation did not use blocks and stratification, and so appears considerably unbalanced. Some levels of the categorical variables might not be relevant to the actual study or are missing; these will be adapted automatically by the software package for the clinical data. Treatment detail variables (such as tumour volumes and doses) have not been finalised and may be updated prior to publication. Further, variables included in the table are the minimum set of variables for the publication, but others might be added.

Patient characteristics and treatment details.

Variable	Overall	Standard	Escalated
n	360	201	159
Age [years] (median [IQR])	60.5 [53.0, 67.0]	60.0 [54.0, 68.0]	61.0 [52.5, 66.0]
Sex = Male/Female (%)	175/185 (48.6/51.4)	94/107 (46.8/53.2)	81/78 (50.9/49.1)
Histology (%)			
Squamous carcinoma	81 (22.5)	47 (23.4)	34 (21.4)
Adeno carcinoma	75 (20.8)	40 (19.9)	35 (22.0)
Adenosquamous carcinoma	111 (30.8)	63 (31.3)	48 (30.2)
NOS	93 (25.8)	51 (25.4)	42 (26.4)
Stage (%)			
IB	54 (15.0)	30 (14.9)	24 (15.1)
IIA	58 (16.1)	33 (16.4)	25 (15.7)
IIB	66 (18.3)	36 (17.9)	30 (18.9)
IIIA	71 (19.7)	40 (19.9)	31 (19.5)

Patient characteristics and treatment details.

Variable	Overall	Standard	Escalated
IIIB	63 (17.5)	33 (16.4)	30 (18.9)
IV	48 (13.3)	29 (14.4)	19 (11.9)
Performance status = 0/1 (%)	184/176 (51.1/48.9)	111/90 (55.2/44.8)	73/86 (45.9/54.1)
FEV1 % of expected (median [IQR])	64.1 [46.1, 82.9]	62.3 [45.7, 79.9]	65.2 [47.4, 85.7]
DLCO % of expected (median [IQR])	55.0 [42.0, 68.0]	55.0 [41.0, 68.0]	55.0 [42.0, 66.0]
Smoking at start RT = No/Yes (%)	237/123 (65.8/34.2)	128/73 (63.7/36.3)	109/50 (68.6/31.4)
No. RT fractions (%)			
<30	1 (0.3)	1 (0.5)	0 (0.0)
30-32	1 (0.3)	1 (0.5)	0 (0.0)
33	358 (99.4)	199 (99.0)	159 (100.0)
No. days during RT (median [IQR])	40.0 [38.0, 42.0]	40.0 [38.0, 42.0]	40.0 [38.0, 42.0]
Mean dose to T [Gy] (median [IQR])	68.5 [65.8, 79.7]	66.0 [65.0, 67.4]	80.2 [78.8, 81.2]
Mean dose to N [Gy] (median [IQR])	68.6 [65.8, 79.5]	66.0 [64.5, 67.4]	80.0 [78.4, 81.7]
Mean lung dose [Gy] (median [IQR])	15.1 [13.7, 16.3]	15.2 [13.7, 16.4]	14.8 [13.9, 16.0]
Volume of T (median [IQR])	10.0 [2.3, 34.6]	9.6 [2.5, 36.9]	10.1 [2.0, 28.1]
Volume of N (median [IQR])	6.6 [1.5, 19.1]	6.4 [1.4, 19.0]	6.8 [1.5, 19.7]
No. Navelbine (median [IQR])	2.0 [2.0, 3.0]	3.0 [2.0, 3.0]	2.0 [2.0, 3.0]
No. Platinum during RT (%)			
1	94 (26.1)	46 (22.9)	48 (30.2)
2	172 (47.8)	99 (49.3)	73 (45.9)
3	94 (26.1)	56 (27.9)	38 (23.9)

Section 6: Analysis

Part 26: Outcome definitions

The current analysis of the NARLAL II trial focuses on the secondary endpoint of early treatment-related toxicity (within 6 months after randomisation). Toxicity has been scored according to CTCAE version 4.0. Toxicity was scored in three ways: 1) prospectively, 2) retrospectively, and 3) based on SAE reports. The prospective toxicity scoring was performed at baseline, weekly during the radiotherapy course and 3 and 6 months after randomisation. The prospectively scored toxicities were:

1) Fatigue
2) Cough
3) Dyspnea
4) Constipation (only during RT)
5) Nausea
6) Vomiting
7) Dysphagia
8) Pain

9) Sensory neuropathy
10) Infection
11) Diarrhoea
12) Skin reaction (only during RT)
13) Ototoxicity (only during RT)
14) Other toxicity

Dysphagia included all esophagitis scores. All prospectively scored values used all six levels of the CTCAE scale (0-5) without any level grouping. Furthermore, patient performance status (according to WHO performance score) was evaluated at all follow-up visits. A patient reached the trial's primary endpoint when a local recurrence was detected. Thus, the data quality of the toxicity scores is reduced significantly after the time of local recurrence. Evaluation of toxicity following recurrence is clinically challenging, as any symptoms may be attributed both to toxicity and to the tumour recurrence at the discretion of the local investigator. Thus, in the current analysis, toxicity after local recurrence will only be reported for grades 4-5.

Pneumonitis within the first 6 months (i.e. for the early toxicity reporting) was scored retrospectively. Peumonitis was scored using CTCAE values; however, levels 0-1 were grouped during the scoring while levels 2, 3, 4, 5 were scored individually. For the retrospective scoring, only the maximum toxicity score and the date of this score were noted, so there is no patient-specific information about the time profile of pneumonitis.

All reported SAE were retrospectively evaluated for radiotherapy-related toxicity and assigned scores according to CTCAE version 4.0. As part of this, it was reviewed whether all prospective scores (the 14 toxicity types listed above) included the reported toxicity of the SAE. If not, the toxicity score from the SAE was added at the nearest time point. In addition to the above-described prospectively scored toxicity types, five additional retrospectively scored SAE groups were utilised for the categorisation of the SAE reporting:

1) Febrile neutropenia
2) Hemoptysis
3) Heart event
4) Emboli
5) Infection (pneumonia-related)

Only grade 3-5 toxicities are reported as SAEs; thus, patients without an SAE report will have grades 0-2. For infection, only the pneumonia-related scores were extracted from the SAE reports (in contrast to the prospectively scored infection that included all types of infections). Similar to the retrospective scored toxicities (febrile neutropenia, hemoptysis, heart event, emboli, and infection (pneumonia-related)), only the maximum grade per toxicity type and the date of this maximum are reported per patient.

All toxicity scores assigned to "other toxicity" were evaluated, and any toxicity deemed belonging to a specified type will be reclassified. After potential reclassification, "other toxicity" will only be reported if some specific "other toxicity" type becomes evident during the analysis. Sensory neuropathy, ototoxicity and remaining "other toxicity" will not be included in the current report, as all are deemed non-related to the RT/Chemo treatment.

Specific toxicities can be challenging to separate clinically. Thus, in addition to the above individual toxicities, three different aggregated groups of toxicity will also be reported based on the maximum toxicity grade within the groups:

1) Lung: cough, dyspnoea, and radiation pneumonitis.
2) Oesophageal: dysphagia, which also includes esophagitis.
3) Gastrointestinal: constipation, diarrhoea, nausea, and vomiting.

The lung group combines prospective-scored toxicity as well as retrospective-scored radiation pneumonitis including SAE-reported pneumonitis. For radiation pneumonitis, only the worst grade is reported, and it will be assigned to the follow-up visit closest to the date of maximum toxicity and evaluated as missing information at all other follow-up visits. The described retrospective assignment of toxicity for the lung group might bias the time profile (see analysis Part 27), but it is likely the best compromise. All statistical tests will not be biased by the temporal toxicity assignment method (see Part 27). The Oesophageal and Gastrointestinal groups contain only prospectively scored toxicities. The oesophageal group is identical to dysphagia described above; thus, it is not a statically independent reporting.

Besides the three aggregated groups, an aggregation will be performed over all toxicities using the max toxicity per patient. As for the lung group above, this aggregation will contain prospective, retrospective and SAE-based toxicity reporting.

Finally, all patients with toxicity grade 4 or 5 occurring after local recurrence, which was not observed before local recurrence and is reported no later than 6 months after randomisation, will be evaluated as Unrelated, Unlikely, Possibly, or Related to the treatment. Those not assessed as Unrelated will be reported.

In summary, the following outcomes will be reported:

1) Prospectively scored toxicity, retrospectively scored pneumonitis, and the five additional retrospectively scored SAE groups, reported separately per type. The retrospectively scored groups will be analysed without any time division, while the prospectively scored toxicities will be reported as
    a. During RT
    b. At follow-up months 3 and 6

    Only toxicity until local recurrence is included
2) Aggregated toxicity per organ system (see aggregation lists above), with the worst toxicity for the organ system at each time point
    a. Lung toxicity
        i. During RT
        ii. At follow-up months 3 and 6
    b. Oesophageal toxicity
        i. During RT
        ii. At follow-up months 3 and 6
    c. Gastrointestinal toxicity
        i. During RT
        ii. At follow-up months 3 and 6

    Only toxicity until local recurrence is included
3) Maximum toxicity grade aggregated over all individual reported toxicities
    The maximum grade is aggregated over toxicity during RT and at follow-up months 3 and 6 together; thus, only one max value per patient
    Only toxicity until local recurrence is included

4) Performance status
    a. During RT
    b. At follow-up months 3 and 6
5) Grade 4 and 5 toxicity after local recurrence and their cause
   All toxicity until 6 months after randomisation is included

Part 27: Analysis methods

For all tables, plots, and analyses, grades 0-1 will be grouped since the difference between levels zero and one is deemed of no clinical relevance. Besides this grouping, no additional grouping will be performed except for the grouping needed for SAE-reported/aggregated toxicity (described above).

Data will be presented as tables and figures. Statistical tests of differences between treatment arms will be related to the tables and are described below.

Tables

The following tables will be used to summarise toxicity data:

1. *Specific treatment-related toxicities (see table 2 - related p-values see below)*
   Number and proportion of patients experiencing each specific type of treatment-related toxicity. For all prospectively scored toxicities the maximum grade per patient is reported separately during and after RT, split on study arms. For retrospectively scored toxicities no division into the two groups during and after RT will be performed.
   The specific toxicity types are: Fatigue, Cough, Dyspnea, Constipation (only during RT), Nausea, Vomiting, Dysphagia, Pain, Infection (prospective, all infections), Diarrhoea, Skin reaction (only during RT), pneumonitis (retrospective), Febrile neutropenia (SAE), Hemoptysis(SAE), Heart(SAE), Emboli(SAE), Infection (SAE, pneumonia-related).
   Number of table-related statistical tests: 27
2. *Aggregated toxicity per organ system (see table 3)*
   Similar to point 1, but for each of the three aggregated toxicities (lung, oesophagus, gastrointestinal). Maximum grade per patient reported separately during and after RT, split on study arms.
   Number of table-related statistical tests: 4 (only four since the oesophagus toxicity group is identical to dyspnea tested above)
3. *Maximum toxicity (layout as table 3, but without the separation between "During RT" and "After RT")*
   Maximum overall observed toxicity grade per patient (for any type of toxicity included in point 1), split on study arms (during and after RT is combined into one total tox score per patient). A "lack of scoring" within the five additional retrospective scored SAE groups (thus, a toxicity score less than 3) will be treated as a grade zero within this aggregation since there is no observed toxicity.
   Number of table-related statistical tests: 1
4. *Performance status*
   The numbers and proportion of patients are reported similarly to the layout of table 2 (point 1). All levels of performance status are reported individually; thus, no aggregation between levels 0 and 1 is performed.
   Number of table-related statistical tests: 2

5. *All grade 4 and 5 toxicity after local recurrences*

   All patients with grade 4 and 5 toxicities after local recurrence, which were not observed before local recurrence and are reported no later than 6 months after randomisation, will be evaluated as Unrelated, Unlikely, Possibly, or Related to the treatment. Except for those assessed as Unrelated, grades 4 and 5 will be tabulated (see example in Table 4 below). The table may not necessarily be included as an actual table in the article but can also be reported within the article's main text or in an appendix.

   There is no related statistical test for this outcome.

Table 2: Example of a table showing specific treatment-related early toxicities. The five retrospectively scored SAE groups and pneumonitis are reported in the "After RT" column but include the toxicity during RT

		During RT				After RT			
Toxicity		Total	Standard	Escalated	p-value	Total	Standard	Escalated	p-value
Fatigue					zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Cough					zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Dyspnea					zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Constipation					zz				
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		NA	NA	NA	
	2	xx (yy%)	xx (yy%)	xx (yy%)		NA	NA	NA	
	3	xx (yy%)	xx (yy%)	xx (yy%)		NA	NA	NA	
	4	xx (yy%)	xx (yy%)	xx (yy%)		NA	NA	NA	
	5	xx (yy%)	xx (yy%)	xx (yy%)		NA	NA	NA	
Nausea					zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Vomiting					zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Dysphagia					zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	

Pain				zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
Infection (all infections)				zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
Diarrhoea				zz				zz
	0-1	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	2	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	3	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	4	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
	5	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	xx (yy%)	
Skin reaction				zz				
	0-1	xx (yy%)	xx (yy%)	xx (yy%)	NA	NA	NA	
	2	xx (yy%)	xx (yy%)	xx (yy%)	NA	NA	NA	
	3	xx (yy%)	xx (yy%)	xx (yy%)	NA	NA	NA	
	4	xx (yy%)	xx (yy%)	xx (yy%)	NA	NA	NA	
	5	xx (yy%)	xx (yy%)	xx (yy%)	NA	NA	NA	
Pneumonitis (retrospective)								zz
	0-1	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	2	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	3	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	4	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	5	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
Febrile neutropenia (SAE)								zz
	0-2	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	3	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	4	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	5	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
Hemoptysis(SAE)								zz
	0-2	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	3	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	4	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	5	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
Heart (SAE)								zz
	0-2	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	3	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	4	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	5	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
Emboli (SAE)								zz
	0-2	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	3	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	4	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	5	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
Infection (SAE, pneumonia-related)								zz
	0-2	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	3	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	4	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	
	5	NA	NA	NA	xx (yy%)	xx (yy%)	xx (yy%)	

Table 3: Example of a table showing early toxicities aggregated across organ systems.

	During RT				After RT			
Toxicity	Total	Standard	Escalated	p-value	Total	Standard	Escalated	p-value
Lung				zz				zz
0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Oesophagus (identical to dysphagia in table 2)				zz				zz
0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
Gastrointestinal				zz				zz
0-1	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
2	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
3	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
4	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	
5	xx (yy%)	xx (yy%)	xx (yy%)		xx (yy%)	xx (yy%)	xx (yy%)	

Table 4: Example of a table showing individual grade 4 and 5 toxicity events after the occurrence of local recurrence and before 6 months after randomisation. The table only includes toxicity scored with a higher grade than the scored grade before the occurrence of local recurrence. Only toxicities which are not deemed unrelated to the treatment are included in the table.

Assigned study arm	Grade	Category	Term	Related to treatment	Days from RT start
Standard treatment	4	Pulmonary	Radiation pneumonitis	Related	40
	4	Haemorrhage	Pulmonary haemorrhage	Possibly	112
	5	Pulmonary	Dyspnoea	Unlikely	84
	Xx	xx	xx	xx	Xx
	Xx	xx	xx	xx	Xx
	Xx	xx	xx	xx	Xx
Dose escalation	4	Vascular	Thrombosis	Unlikely	55
	5	GI	Tracheo-oesophageal fistula	Related	109
	5	Pulmonary	Radiation pneumonitis	Related	51
	Xx	xx	xx	xx	Xx
	Xx	xx	xx	xx	xx

Figures

The tables in points 1,2, and 4 aggregate the information in two groups during and after RT. Figures related to each of the above tables will be provided to illustrate the temporal behaviour of the toxicity. An example of such a sub-figure is shown in Figure 4, which at each time point shows the proportion of patients who scored

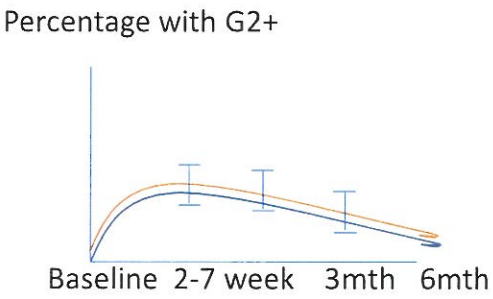
a) grade 2 or above
b) grade 3 or above

Time profiles will only be made for the prospectively scored toxicities and not for the retrospectively scored toxicities (the five additional retrospectively scored SAE groups and retrospectively scored pneumonitis) since their related temporal information is limited. However, for the aggregated lung group, which contains retrospectively scored pneumonitis, the time profile is plotted since, for the aggregation purpose, the pneumonitis scores are attempted to be placed on the relevant follow-up time (described above).

For the performance status (point 4 within the table list), the proportion of patients with performance 2+ and 3+ will be plotted.

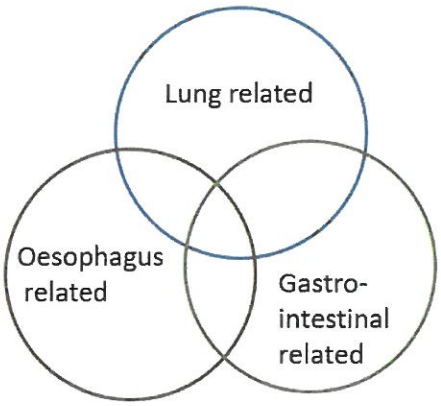
Note that the above-mentioned figures show the ratio between the reported toxicity of the specified grade and all evaluated patients at that specific time. This approach can be biased if there is a considerable imbalance in the number of patients not evaluated at individual time points (e.g., due to large differences in early recurrence rates). In that case, further sensitivity analyses may have to be considered.

Figure 4.: Example of the percentage of patients with grade 2 or more dysphagia (the x-axis will also include all the toxicity values for weeks 2 to 7).



Finally, a Venn diagram (see Figure 5) will be used to illustrate potential correlations between the three aggregated outcomes in Part 26 (see table definition point 2). This will show the overlap in patient populations who experienced at least grade 3 (grade 3+) toxicity for each of the three organ systems examined at any time point within the evaluation period for early toxicity.

Figure 5: Venn diagram illustrating patients with more than one event type of grade 3 or more. Event types are related to the three organ-aggregated toxicities: lung-related, oesophagus-related and gastrointestinal-related.



Comparative analysis of toxicity in treatment arms

Mixed effect ordinal logistic regression models will be used to evaluate each outcome (1-4) described in Part 26. Toxicity during and after RT will be evaluated separately. Each toxicity observation (weeks during RT or months 3 and 6) will be treated as individual observations for the given patient. The toxicity levels in the outcome model will be as shown in Tables 2 and 3. The predictors in the model will be baseline toxicity and treatment arms as fixed effects, with individual patients as a random effect (for outcomes in which all patients only provide one observation per patient, the random effect will be omitted). For toxicity after RT, Durvalumab treatment will be added as a fixed effect, and the status variable for Durvalumab will only be true if Durvalumab treatment is started before the specific toxicity scoring (R model description see footnote [B]). Except for the separate data analysis during and after RT, there is no test of temporal toxicity dependence. The significance level for each outcome under consideration will be based on the p-value related to the regression constant of the treatment arm in the outcome-specific regression model. With the above-stated tests, there will be 34 statical tests (including performance status and excluding oesophagus, which only includes dysphagia). The number might be larger if any specific "other toxicity" is analysed. All p-values related to toxicity differences between treatment arms will be included in the multiple-testing approach described in Part 17. There will be no multiple testing correction of the p-values related to Durvalumab. The strategy related to missing baseline data is described in Part 28. All non-adjusted p-values related to the treatment arm will be reported in Tables 2 and 3 (see above). Those that are significant after multiple testing adjustment will be visually indicated, e.g. by an added star.

Note specifically that

- The proposed approach (mixed effect ordinal logistic regression models) can be considered as an extension to simple tests of differences in tabulated scores, such as a χ^2 test. However, this approach takes multiple toxicity evaluations per patient into account, while being able to account for missing data (missing toxicity evaluations) and as well as other fixed effects (such as baseline scores and Durvalumab).
- This approach represents a slight deviation from the approach specified in the original trial protocol (which stated that the toxicity would be analysed using non-parametric methods). However, this is a much more robust approach, while not deviating from the intentions of the original plan.

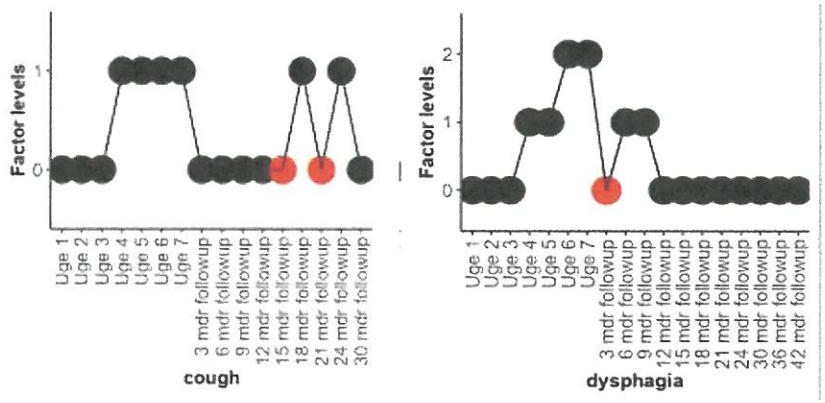
[B] MixedModelResult <- ordinal::clmm(Toxicity ~ BaselineTox+TreatmentArm+Durvalumab+(1|PtId), data = data)

Part 28: Missing data

Part of the data curation done before analysis is plotting all toxicity measured per patient and toxicity type. See the example in Figure 6. All missing data are shown in the figure as red dots (plotted with the value corresponding to the most frequent observation). For a patient with missing toxicity data for a specific toxicity and with at least one score of 2 or above at other follow-up visits, the electronic patient journal will be consulted for information on the missing toxicity scorings, and the physician will check if an SAE or notes in the patient journal can be used to assess the toxicity score. If not, the specific toxicity will be missing. This procedure should minimise the risk of missing severe toxicity scores.

During analysis, toxicity for a patient with missing toxicity data and for which all other scores for the given toxicity are evaluated as 0 or 1 will be interpreted as toxicity of zero, if there is documentation that the patient did show up for the given follow-up visit (there are toxicity score for other types of toxicity). The decision to perform this automatic imputation was made since unscored data will most likely represent no toxicity, as unscored data either stem from missed input of score (not very likely in case of observed toxicity) or the expectation that no scoring is identical to zero.

Figure 6. Black dots illustrate prospectively scored toxicities during patient visits. Red dots illustrate missing scores. Left: This patient did not show up for consultation at 15 months and the point will be missing. The patient showed up at 21 months and the point will be allocated to 0 as only grade 0 or 1 coughing has been scored. Right: The patient showed up at the consultation at 3 months. The patient journal and SAEs will be checked for information on dysphagia.



Baseline toxicity is used as a toxicity predictor in the planned regression analysis. Consequently, if baseline toxicity is missing for selected patients, multiple data imputation for the baseline value will be performed. Each imputation of the baseline value will be sampled from the distribution of baseline values for the specific toxicity of the contributing institution. If an institution has provided less than 20 patients, the imputation will be based on the distribution of the entire trial cohort. The multiple imputations and related regression analyses will be performed 20 times (using a new baseline data imputation each time), and the reported p-value will be the median of the 20 calculated p-values.

There will be no attempt to impute the outcome toxicity. This means that the obtained result is based on the available toxicity scores, which, as mentioned previously, could bias the result if there is an imbalance between the patient showing up for follow-up between the two treatment arms. In case of a large toxicity difference between the treatment arms, the frequency of reported toxicities for each arm will be calculated to evaluate the potential impact of such an imbalance.

Part 29: Additional Analyses

The investigator group has decided not to include any analysis of a potential difference in toxicity reporting between the participating institutions. All institutions have evaluated toxicity according to a predefined scale (CTCAE 4.0), and thus any differences should ideally represent centre differences in normal tissue irradiation (which is beyond the scope of the current report). However, centre differences might be present due to other effects (such as differences in clinical evaluations or in patient demographics); thus, an analysis of the association between toxicity scores, individual centres, and normal tissue irradiation may be conducted in the future.

Part 30: Harms

The entirety of this SAP is related to toxicity differences between the two treatment arms. Besides potential toxicity differences, there are no additional harms which should be addressed as part of the current SAP.

Part 31: Statistical Software

In line with the article for the primary endpoint of NARLAL II, a statistical package written in R will be developed. The package will be publically available and accessible at https://github.com/oncology-ouh/Narlal2.

The final analysis will be conducted in R and based on the package, which will also facilitate easy export to other statistical software systems.

Part 32: References

1) Peter C. O'Brien TRF. A Multiple Testing Procedure for Clinical Trials. Biometrics. 1979;35:549-56.
2) Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57:289-300
3) Lokke A, Marott JL, Mortensen J, Nordestgaard BG, Dahl M, Lange P. New Danish reference values for spirometry Clin Respir J. 2013;7(2):153-67.