Statistical Analysis Plan

Principal Investigator:	Jennifer S. Temel, MD
Protocol Number and Title:	17-471 Randomized Trial of Stepped Palliative Care versus Early Integrated Palliative Care in Patients with Advanced Lung Cancer
Protocol Version and Date:	Version 12.0, 03 February 2023
Author(s):	Nora Horick, MS, Senior Biostatistician Kedie Pintro, MS, Biostatistician Dustin Rabideau, PhD, Faculty Biostatistician
SAP Version:	Version 1.0
SAP Version Date:	October 6, 2023

Revision History

Version #	Date	Document Owner	Revision Summary
Version 1.0	06-Oct-2023	Nora Horick	Initial Release Version

Signature

I confirm that I have reviewed this document and agree with the content.

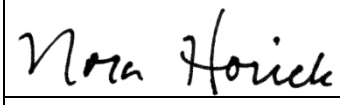
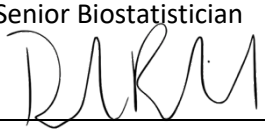
APPROVALS	
	10/13/2023
Nora Horick, MS Senior Biostatistician	**Date**
	10/13/2023
Dustin Rabideau, PhD Faculty Biostatistician	**Date**
	10/13/2023
Jennifer S. Temel, MD Principal Investigator	**Date**

Table of Contents

1. Overview

Early and longitudinal involvement of palliative care (PC) in the outpatient management of patients with advanced cancer improves patient-reported and end of life (EOL) care outcomes. While recommended by national organizations as the standard of care, this early integrated care model utilizes substantial PC resources, which has limited its dissemination across care settings. The STEP trial is a randomized trial of stepped PC versus early integrated PC in patients with advanced lung cancer. By demonstrating the non-inferiority of a stepped PC model to early integrated PC, we seek to define a role for this more accessible, scalable, and patient-centered approach to PC.

2. Study Design

STEP is an unblinded, multi-center, non-inferiority randomized trial of stepped PC versus early integrated PC in patients with advanced lung cancer. 510 patients with advanced lung cancer receiving their care at Massachusetts General Hospital, Duke Cancer Center, or University of Pennsylvania Abramson Cancer Center were enrolled. Patients were randomized in 1:1 fashion and stratified by study site and underlying diagnosis. As patients with NSCLC have a significantly better prognosis than those with small cell lung cancer or mesothelioma, stratification by the underlying diagnosis was used to ensure adequate and balanced representation between the two study groups.

The primary hypothesis is that patients receiving stepped PC experience non-inferior QOL at 24 weeks from enrollment compared to patients receiving early integrated PC. Key secondary hypotheses include (1) demonstrating non-inferiority of stepped PC in the rate by which patients communicate their EOL care preferences to their clinicians and with respect to patients' length of stay in hospice, and (2) showing that stepped PC utilizes fewer PC resources than early integrated PC.

3. Patient Population

The patient eligibility criteria mirror those of our prior early PC studies in this patient population.

Inclusion Criteria:

1. Diagnosed with advanced NSCLC, small cell lung cancer, or mesothelioma, being treated with non-curative intent, and informed of advanced disease within the prior twelve weeks
2. Eastern Cooperative Oncology Group (ECOG) Performance Status from 0 (asymptomatic) to 2 (symptomatic and in bed <50% of the day)
3. The ability to read and respond to questions in English or Spanish
4. Primary cancer care at one of the three participating sites
5. Age > 18 years

Exclusion Criteria:

Patients were excluded if:

1. They were already receiving outpatient PC or hospice services since diagnosis of advanced NSCLC, small cell lung cancer, or mesothelioma
2. They have cognitive or psychiatric conditions as determined by the treating oncologist to prohibit study consent or participation

4. Randomization

Within 2 weeks of providing informed consent, patients completed baseline demographic and study questionnaires. Once baseline measures were completed, patients are randomized in a 1:1 fashion, stratified by study site (MGH vs Duke vs Penn) and cancer diagnosis (NSCLC vs SCLC and mesothelioma) using a computer-generated randomization schema. Both the participating study clinicians and patients were aware of the study group assignments, since the frequency and timing of intervention visits precluded blinding PC clinicians or patients.

5. Intervention Delivery

Early Integrated PC
Patients randomized to early integrated PC were scheduled to meet with a PC clinician within 4 weeks of enrollment and at least every 4 weeks throughout their disease course. If a patient missed a scheduled visit or is unable to be scheduled within 4 weeks of their last PC visit, a PC clinician attempted to call them by telephone to maintain contact at least every 4 weeks and rescheduled the visit as soon as possible. The inpatient PC team followed patients who were admitted to a study site hospital.

Stepped PC
Patients randomized to stepped PC were scheduled for an initial visit with a PC clinician within 4 weeks of enrolment. During step 1, further visits with a PC clinician were scheduled at clinically significant points in the patient's illness, including within 4 weeks of (1) a change in cancer treatment (due to either progression or toxicity) or (2) hospital discharge. After each visit, the PC clinician communicated with the oncology clinician(s) either by telephone, email or in person. If a patient missed a scheduled visit or was unable to be scheduled for a PC visit, the PC clinician attempted to contact them by telephone and rescheduled the visit as soon as possible. Patients assigned to stepped PC completed the FACT-L every 6 weeks during the first 18 months of study participation. Those whose scores decreased by ≥10 points from baseline were 'stepped up' to step 2 and followed the same protocol as those randomized to the early integrated PC arm.

All study participants in both groups surviving greater than 18 months from enrolment were permitted to decrease the frequency of PC visits as per their preference and the discretion of their PC and oncology clinicians.

6. Outcome Measures

Outcome data included self-report questionnaires collected prior to randomization (baseline) and then again at weeks 12, 24, 36 and 48 (with a ±2-week window) (see Table 1), as well as information collected from the EHR.

Table 1. Study Questionnaires

Self-report measure	Baseline	Every 6 weeks	Weeks 12, 24, 36 & 48	Scoring
Demographics	X			n/a
SCQ	X			Range: 0-45 (higher is worse medical condition)
FACT-L	X	X[1]	X	Range: 0-136 (higher is better QOL)
PHQ-9	X		X	Range: 0-27 (higher is worse depression)
PTPQ	X		X	n/a

Brief Cope	X		X	Range: 2-8 for each facet (higher is more engagement in coping style)
EQ-5D	X		X	VAS Range: 0-100 (higher is better health state)
Support Service Utilization Item			X[2]	

[1] Step 1 patients completed FACT-L every 6 weeks for up to 18 months from enrollment.
[2] Collected only at week 24.
EQ-5D, EuroQol—5 Dimension; FACT-L, Functional Assessment of Cancer Therapy-Lung; PHQ-9, Patient Health Questionnaire-9; PTPQ, Prognosis and Treatment Perceptions Questionnaire; SCQ, Self-administered Comorbidity Questionnaire.

The primary outcome is patient-reported quality of life at 24 weeks from enrollment, as measured by the FACT-L.

The secondary outcomes are:

1. Patient report of patient-clinician communication about EOL care preference using the following item on the PTPQ: "Have you and your doctors discussed any particular wishes you have about the care you would want to receive if you were dying?" analyzed as "yes" versus "no." The patient's final assessment will be used.
2. Length of stay on hospice (in days) among patients who die during the study, as documented in the EHR.
3. PC resource utilization; specifically, the number of outpatient PC visits per patient during the study, as documented in the EHR.

The exploratory outcomes are:

1. Rate of change in quality of life from baseline to week 48, as measured by the FACT-L.
2. Cost-effectiveness as assessed by costs relative to quality adjusted life years. ***Note:*** *Plans for the analysis of cost-effectiveness are not described in this SAP.*
3. Patient-reported coping strategies at week 24, as measured by eight subscales of the Brief Cope, including subscales measuring emotional support, positive reframing, active coping, acceptance, self-blame, denial, spiritual coping, and behavioral disengagement, as well as higher-order factors reflecting active coping and avoidant coping).
4. Patient-reported prognostic understanding at week 24, based on relevant items on the PTPQ, including:
    a. Patient's primary goal of current cancer care. Responses will be dichotomized based on whether the patient selected "to cure my cancer" versus any of the other options.
    b. Patient's endorsement of the statement "My cancer is curable," analyzed as "yes" versus "no."
5. Patient-reported depression symptoms at week 24, as assessed by the PHQ-9.
6. Healthcare utilization at the end of life (i.e., the 30 days prior to death) among patients who die during the study, including:
    a. Number of emergency department visits
    b. Number of hospitalizations
    c. Chemotherapy administration

7. Sample Size

Sample size calculations were performed to detect a non-inferiority margin of 4.5 points on the FACT-L at week 24, assuming a standard deviation of 17.5 points and a between-group difference of 0 points. 188 patients per group were required to achieve 80% power with a one-sided significance level of 0.05. To account for an anticipated 36% rate of missing data at week 24 due to loss-to-follow-up, withdrawal, or death, enrollment of 255 per group (510 total) was planned.

8. General Statistical Considerations

Statistical Software

All statistical analyses will be performed using SAS (SAS Institute Inc., NC, USA) and R (R Foundation for Statistical Computing, Vienna, Austria).

Analysis Population

The primary analyses will be conducted according to the intention-to-treat principle. All randomized patients will be included and analyzed in the group to which they were initially randomized, regardless of intervention adherence. If there is considerable non-adherence to the randomized intervention, we will conduct additional *post hoc* per-protocol analyses (e.g., using causal inference g-methods, such as inverse probability weighting) to assess whether study results are sensitive to non-adherence to the assigned intervention group.

Patient Disposition

The flow of patients through the study will be demonstrated using a flow diagram, consistent with the Consolidated Standards of Reporting Trials (CONSORT) statement. This diagram will display number of patients assessed for eligibility, the number of patients who enrolled and randomized in the study and the number of patients who were excluded or otherwise not enrolled. For enrolled patients, the study group allocation will be displayed, along with the number of patients who complete follow-up assessments at each time point. The number of patients in each group who were lost to follow-up or otherwise excluded from analysis will be displayed.

Baseline Characteristics

Baseline characteristics including demographic information, smoking status, cancer type, comorbidities, and baseline PROs score will be summarized by intervention group using descriptive statistics and visual displays. Descriptive statistics for continuous variables will include the number of subjects, mean, standard deviation, median, first and third quartiles, and minimum and maximum values for the observed value. Frequencies and percentages will be calculated for categorical variables.

Adjustment of the Intervention Effect

The primary analyses for all outcomes will be adjusted for the randomization stratification variables: study site (MGH vs Duke vs Penn) and cancer diagnosis (NSCLC vs SCLC and mesothelioma). Small strata will be combined as needed to ensure numeric stability of model estimates. If needed, sensitivity analyses will additionally adjust for baseline characteristics that are imbalanced between treatment groups. Imbalanced characteristics will be identified *a priori* by examining the distributions of baseline characteristics known to be associated with the primary and secondary outcomes by intervention group. Characteristics that are imbalanced to an extent that is considered clinically meaningful will be included as adjustment covariates in addition to study site and cancer diagnosis.

Significance Level and Multiplicity Adjustment

Statistical significance testing for non-inferiority of the primary outcome will be one-sided with a type I error rate of 0.05. Testing of secondary outcomes will be one-sided for outcomes being tested for non-inferiority and two-sided for outcomes being tested for superiority. The Benjamini-Hochberg false discovery rate (FDR) control approach will be used to interpret results of significance tests of secondary outcomes with an FDR of 0.15. Analyses of exploratory outcomes will not be adjusted for multiple comparisons, and presented results will emphasize estimates and confidence intervals.

Scoring of Patient-Reported Outcomes

Total and subscale scores for patient-reported outcomes including FACT-L, PHQ-9, and Brief Cope will be calculated for each patient/timepoint using published scoring algorithms. When scoring instructions specifically address how missing item responses should be handled in the score calculation, these instructions will be followed. If no specific guidance for handling missing item responses is provided, subscale scores (if applicable) or total scores (for unidimensional scales) will be calculated via single imputation of the mean of observed subscale values if >50% of the subscale items are non-missing. For scales with total scores derived from the sum or average of two or more subscales, the total score will not be calculated unless all subscale scores can be calculated.

Primary analyses of FACT-L scores will utilize the standard published scoring method. Secondary analyses will use an alternative scoring method for surveys that do not meet the above criteria for calculating the total score—i.e., those that have 50% or fewer non-missing responses on one or more subscales. The alternative scoring method allows calculation of the total score if >50% of all items (across all four subscales) are non-missing.

Missing Data

Primary analyses will include available data without imputation of missing data. The characteristics of patients who complete versus do not the week 24 survey will be compared descriptively. Sensitivity analyses will explore how different assumptions about the missing data mechanisms affect estimated outcomes. These sensitivity analyses may include multiple imputation, terminal decline joint modeling, or partially conditional models, which provide estimates of the mean conditional on being alive and observed at each timepoint.

9. Statistical Analysis

Primary Outcome

The primary outcome is patient-reported quality of life at week 24, as measured by the FACT-L. The difference in week 24 means between intervention groups will be estimated using a linear regression model with group assignment (stepped PC vs early integrated PC) and baseline FACT-L score, as well as randomization stratification factors, as main effects. Using the pre-specified non-inferiority margin of 4.5 points, non-inferiority of stepped PC will be established if the lower limit of the 90% confidence interval (CI) for the estimated difference in means is greater than -4.5. The corresponding p-value for non-inferiority will be calculated using a one-sided test of the estimated difference in means against a null value of -4.5 with a significance level of 0.05. Model-based estimates of the mean with 95% CI in each group, the difference in means between groups and its 90% CI, and the p-value will be reported.

To accompany the approach described above, we will also use a linear mixed effects regression model to estimate the difference in week 24 means between intervention groups. This model will utilize FACT-L scores at baseline, week 12, and week 24, with estimation via maximum likelihood. The model will include fixed effects for group assignment (stepped PC vs early integrated PC), time from baseline (in weeks), time-by-group interaction, and randomization stratification factors, as well as random intercepts for each patient. A contrast will be used to estimate the difference in means at week 24, and this estimate will be evaluated for non-inferiority in the same manner described above. This will be reported as a sensitivity analysis for evaluating non-inferiority for the primary outcome.

Secondary Outcomes

The difference between groups in the rate of patient-clinician communication about EOL care preference at the final follow-up assessment will be assessed using a generalized linear regression model specified with an identity link function and binomial response probability distribution. The model will include group assignment (stepped PC vs early integrated PC) and randomization stratification factors as main effects. Using the pre-specified non-inferiority margin of 10%, the nominal p-value for non-inferiority will be calculated using a one-sided test of the estimated difference in rates against a null value of -10%. Non-inferiority of stepped PC will be established if the nominal p-value is significant after application of the multiplicity adjustment procedure for secondary outcomes. Model-based estimates of the rate with 95% CI in each group, the difference in rates between groups and its 90% CI, the nominal p-value, and the multiplicity-adjusted p-value will be reported.

The difference between groups in the mean length of stay in hospice among patients who die during the study will be assessed using a linear regression model with main effects for group assignment (stepped PC vs early integrated PC) and randomization stratification factors. Using the pre-specified non-inferiority margin of 7 days, the nominal p-value for non-inferiority will be calculated using a one-sided test of the estimated difference in rates against a null value of -7. Non-inferiority of stepped PC will be established if the nominal p-value is significant after application of the multiplicity adjustment procedure for secondary outcomes. Model-based estimates of the mean with 95% CI in each group, the difference in means between groups and its 90% CI, the nominal p-value, and the multiplicity-adjusted p-value will be reported.

The difference between groups in the mean number of outpatient PC visits per patient will be assessed using a linear regression model with main effects for group assignment (stepped PC vs early integrated PC) and randomization stratification factors. Superiority of stepped PC will be established if the nominal p-value for the difference in means (against a null value of 0) is significant in favor of stepped PC after application of the multiplicity adjustment procedure for secondary outcomes. Model-based estimates of the mean with 95% CI in each group, the difference in means between groups and its 95% CI, and the nominal p-value, and the multiplicity-adjusted p-value will be reported.

Exploratory Outcomes

The rate of change in quality of life from baseline to week 48 (assessed by the FACT-L) will be compared between groups using a linear mixed effects regression model. This model will utilize FACT-L scores at each timepoint, with estimation via maximum likelihood. The model will include fixed effects for group assignment (stepped PC vs early integrated PC), time from baseline (in weeks), time-by-group interaction, and randomization stratification factors, as well as random intercepts for each patient.

Model-based estimates of the slope with 95% CI in each group and the difference in slopes between groups and its 95% CI will be reported.

The difference in means between intervention groups in patient-reported depression symptoms (assessed by PHQ-9) and coping strategies (assessed by subscales and higher-order factors of Brief Cope) at week 24, will be estimated for each outcome using a linear regression model with group assignment (stepped PC vs early integrated PC) and baseline score, as well as randomization stratification factors, as main effects. Model-based estimates of the mean with 95% CI in each group and the difference in means between groups and its 95% CI will be reported.

Rates of prognostic understanding at week 24, based on PTPQ items eliciting the patient's goal of cancer care ("to cure my cancer" vs any other option) and the patient's assessment of curability ("yes" vs "no") will be compared using generalized linear models specified with an identity link function and binomial response probability distribution. Model-based estimates of the rate with 95% CI in each group and the difference in rates between groups and its 95% CI will be reported.

Multiple metrics of healthcare utilization at the end of life will be compared between intervention groups among patients who die during the study. The number of emergency department visits and number of hospitalizations at the end of life will be compared using linear regression models or count regression models (e.g., Poisson, negative binomial) if the linear regression distributional assumptions are violated. The rates of chemotherapy administration (any versus none) at the end of life will be compared using generalized linear models specified with an identity link function and binomial response probability distribution. Additionally, the occurrence of any emergency department visit or any hospitalization at the end of life may be compared using these binomial models. Model-based estimates of the mean or rate with 95% CI in each group and the difference in means or rates between groups and its 95% CI will be reported.

This addendum to the STEP SAP includes detailed clarifications and required modifications that were made to the analysis plan after the SAP had been finalized and signed.

Index	Date	**Topic:** Description
1	11/21/2023	**Change in scoring method used for Self-Administered Comorbidity Questionnaire (SCQ):** This item clarifies how the SCQ scores were calculated such that the range for scores is 0-36. The SCQ includes 13 conditions with 3 yes/no questions (scored 1/0) about each condition. In addition, there are fields for 2 "other" (write in) conditions, each with 3 yes/no questions. The score for each condition ranges from 0-3, and scores for each condition are summed to compute the total SCQ score. We did not include the "other" conditions in the scoring, and we combined responses for osteoarthritis and rheumatoid arthritis into a single "arthritis" condition based on the approach used in the primary SCQ manuscript. Thus, the SCQ scores could range from 0-36.
2	12/21/2023	**Inclusion of week 12 assessments for prognostic understanding (exploratory outcome):** This item clarifies that we will use the final post-baseline prognostic understanding assessment for exploratory outcome #4 (i.e., we will use week 12 prognostic understanding if week 24 is missing): 1. *Patient-reported prognostic understanding at week 24* ***(or week 12 if week 24 is missing)****, based on relevant items on the PTPQ, including:* a. *Patient's primary goal of current cancer care. Responses will be dichotomized based on whether the patient selected "to cure my cancer" versus any of the other options.* b. *Patient's endorsement of the statement "My cancer is curable," analyzed as "yes" versus "no."*
3	1/5/2024	**Week 24 analysis time point for PC resource utilization (secondary outcome):** To align with the primary outcome (evaluated at 24 weeks), we will evaluate secondary outcome #3 in the SAP through week 24: 3. *PC resource utilization; specifically, the number of outpatient PC visits per patient* ***through week 24****, as documented in the EHR.*