TOPaZ

Treatment of Osteogenesis Imperfecta with Parathyroid hormone and Zoledronic acid

Statistical Analysis Plan

CONFIDENTIAL

Version No	1.0
Date	22nd June 2018
Author(s)	Catriona Keerie (to 10 Nov 2017); Christopher Weir (from 10 Nov 2017)
CI Name	Professor Stuart Ralston
CI Email address	stuart.ralston@ed.ac.uk

Signatures	
Trial Statistician: Professor Christopher Weir	**Date:** 25/07/2018
Chief Investigator: Professor Stuart Ralston	**Date:** 25TH JUNE 2018

Document Control		
Version No	**Date**	**Details/reason for update**
DRAFT 1.0	7th April 2017	New document
DRAFT 1.0_27Apr2017	27th April 2017	Revisions following internal review
DRAFT 1.0_03Jul2017	3rd July 2017	Revisions following internal review
DRAFT 1.0_07Aug2017	7th August 2017	Revisions following internal review
DRAFT 1.0_01Nov2017	1st November 2017	Revisions following internal review
DRAFT 1.0_23Nov2017	23rd November 2017	Further internal review
1.0	22nd June 2018	Add propensity scoring details; clarification of adherence assessment; correction of molecular diagnosis groupings; clarification of menarche/menopause data in section 4.2; accept changes for sign-off

Contents

List of Abbreviations

Abbreviation	Full name
AE	Adverse Event
AR	Adverse Reaction
BMD	Bone Mineral Density
BMI	Body Mass Index
BPI	Brief Pain Index
CACE	Complier-average causal effect
CI	Confidence Interval
COLIA1	Collagen type IA1 gene
COLIA2	Collagen type IA2 gene
CONSORT	CONsolidated Standards of Reporting Trials
DMC	Data Monitoring Committee
ECTU	Edinburgh Clinical Trials Unit
EDTA	Ethylene diamine tetraacetic acid
eGFR	Estimated Glomerular Filtration Rate
EME	Efficacy and Medicines Evaluation
EQ-5D	Euroqol 5D measure
HAQ	Health Assessment Questionnaire
IQR	Inter quartile range
ITT	Intention-to-treat
MAR	Missing at random
Max	Maximum
Min	Minimum
MNAR	Missing not at random
N	Number of patients with an observation
OI	Osteogenesis Imperfecta
PP	Per-protocol
PSQI	Pittsburgh Sleep Quality Index
PTH	Parathyroid Hormone
SAE	Serious Adverse Event
SAR	Serious Adverse Reaction
SAP	Statistical Analysis Plan
SD	Standard Deviation
SF-36	Short Form Health Survey
SUSAR	Suspected Unexpected Serious Adverse Reaction
TPTD	Teriparatide
UAR	Unexpected Adverse Reaction
ZA	Zoledronic acid

1. Introduction

This document details the criteria to be used for the definition of the analysis populations and the statistical methodology for analysis for the TOPaZ trial, a randomised controlled trial of teriparatide (TPTD) followed by zoledronic acid (ZA) versus standard care to prevent fractures in adults with Osteogenesis Imperfecta (OI).

This document has been compiled according to the Edinburgh Clinical Trials Unit (ECTU) standard operating procedure (SOP) "Statistical Analysis Plans" ECTU_ST_04 and has been written based on information contained in the study protocol version 7.0, dated 13 February 2018.

TOPaZ is a multicentre, randomised, parallel-group open label trial. Treatment allocation is a 1:1 ratio. Patients will be randomised to either TPTD followed by ZA or to standard care. The aim is to recruit 380 patients with 190 per treatment arm.

2. Statistical Methods section from the protocol

Primary outcome
The primary outcome will be the proportion of participants experiencing a clinical fracture validated by x-ray or other imaging. The main analysis of the primary outcome will summarise time to first fracture by treatment group using Kaplan-Meier survival curves, the groups being compared using the log-rank test stratified by the minimisation variables.
We will review accumulation of fractures during the study and review the situation at 48 months or when 149 patients have suffered a clinical fracture. This will be done by the blinded trial statistician and the DMC will then be asked to make a recommendation to the TSC on continuation (or termination) of the trial.

A secondary analysis will use binary logistic regression, with treatment group (active vs. standard care) and the minimisation variables (fracture in last 2 years, OI clinical subtype, gender, age, BMD group and bisphosphonate use at baseline or in 2 years prior to randomisation) as the independent variables. The effect of randomised treatment will be measured by the odds ratio (and 95% confidence interval) for TPTD/ZA vs. standard care. While every effort will be made to obtain complete follow-up data on all patients, it is recognised that in the OI population some study participants will be lost to follow-up. A sensitivity analysis in which missing data are imputed will be developed according to the principles outlined in (20), namely to develop an understanding for the reasons for loss to follow-up, define the primary set of assumptions about the missing data mechanism on this basis, conduct a statistically valid analysis under these assumptions and explore the robustness of the conclusions in further sensitivity analyses that capture departures from the primary missing data assumptions.

Secondary outcomes
The number of fractures (patient reported and imaging validated) will be analysed using a Poisson regression model, adjusting for the minimisation variables and, if required, including an overdispersion parameter.
Changes in BMD, bone pain, EQ5D, HAQ, SF36, PSQI will be analysed using analysis of covariance (ANCOVA). In each case the model will adjust for the baseline value and the minimisation variables.

Mechanistic Study
The mechanistic objective will be addressed in two stages. First, descriptive statistics of fracture rate will be summarised by treatment group for clinical subtype of OI, baseline BMD, gender and

molecular diagnosis. This will be used to inform a subsequent individual patient data meta-analysis combining the data from this trial and the TPTD and standard care groups from the trial led by co-applicant Professor Bente Langdahl which has started recruitment in Scandinavia (EudraCT 2011-002811-27) and by sourcing data from the trial previously reported by Orwoll in which TPTD was compared with placebo in patients with OI (9). These analyses will include a fixed effect for trial and will formally test, in a separate model for each baseline variable, for an interaction between the baseline variable and the effect of TPTD (versus standard care) on fracture rate. In further pooled analyses data from the standard care groups in both trials will be combined to estimate the association between each baseline variable and fracture risk in patients receiving standard care.

3. Overall Statistical Principles

3.1 SAP objectives

The objective of this SAP is to describe the statistical analyses contributing to the final report and publication(s) of the TOPaZ study. All analyses detailed in the study protocol are addressed with the exception of the second stage of the mechanistic study (individual patient data meta-analysis). This will be documented in a separate IPD meta-analysis plan.

3.2 General principles

In general terms, categorical data will be presented using counts and percentages, whilst continuous variables will be presented using the mean, median, standard deviation (SD), minimum, maximum, inter quartile range (IQR) and number of patients with an observation (n), using the format presented in the examples 1 and 2 below. Data will be split by timepoint where applicable.

Example 1. Data presentation for categorical data

Parameter	Timepoint / visit	Statistic/ category	**TPTD** **N=xx**	**Standard care** **N=xx**	**Overall** **N=xx**
Parameter A	Timepoint x	Category 1	xx (%)	xx (%)	xx (%)
		...	...	...	...
		Category n	xx (%)	xx (%)	xx (%)

Example 2.Data presentation for continuous data

Parameter	Timepoint / visit	Statistic/ category	**TPTD** **N=xx**	**Standard care** **N=xx**	**Overall** **N=xx**
Parameter	Timepoint x	Mean (SD)	xx	xx	xx
		Median	xx	xx	xx
		Q1, Q3	xx, xx	xx, xx	xx, xx
		Min, Max	xx, xx	xx, xx	xx, xx
		n	xx	xx	xx

All applicable statistical tests will be 2-sided and will be performed using a 5% significance level. 95% (2-sided) confidence intervals (CIs) will be presented. All analyses are testing superiority, rather than equivalence or non-inferiority.

Where there is missing data for an outcome variable, in the first instance, those records will be removed from any formal statistical analysis relating to that outcome variable, unless otherwise specified. In tabulations, numbers of missing observations will be provided, but percentages will not include them.

Distributional assumptions underlying the statistical analyses will be assessed by visual inspection of residual plots. Normality will be examined by normal probability plots. If the distributional assumptions for the parametric approach are not satisfied, further data transformation (to alleviate substantial skewness (i.e. normalizing) or to stabilise the variance), or other suitable methods will be considered. This will be documented in the statistical results report together with the reasoning supporting the action taken, if applicable.

All analyses and data manipulations will be carried out using SAS [1].

3.3 Analysis Populations

The analysis populations defined for the statistical reporting of the trial are as follows:

Intention to treat (ITT)	All randomised patients, analysed according to randomised treatment.
Safety population	The safety population will include all patients who were randomised. Patients will be summarised according to treatment received.

A per-protocol population has been considered, but not included for this study since differing patterns of dropouts and departure from randomised treatment between the treatment groups is likely to result in a biased estimate of the treatment effect [3].

By contrast, the complier-average causal effect (CACE) analysis described in Section 4.6 does not depend on the assumption of comparability between groups and is designed to correct for this bias.

4. List of Analyses

4.1 Recruitment and retention

The date of first and last patient randomised, the number of patients eligible, screened, randomised, and the number of centres that recruited patients, will be reported.

A CONSORT flow chart will be provided by the TOPaZ Trial Manager.

This will present the number (%) of patients, split by treatment group, as follows:

- Consented
- Receiving at least one dose of study medication
- Completing the trial
- Discontinuing from the trial

In addition, reasons for non-inclusion in the study (prior to randomisation) will be categorised.

The number of patients discontinued early from the study will be summarised by reason for withdrawal and by treatment group.

4.2 Baseline characteristics (ITT)

Baseline patient characteristics will be summarised by allocated treatment and overall for the ITT population as follows:

- Demographics
  - Age (years)
  - Gender (male/female)
  - For female participants only:
    - Current childbearing potential (Yes/No)
    - Age of menarche (years)
    - Post-menopausal (Yes/No)
    - Age of menopause (years)

- Food frequency questionnaire
  - Total dietary calcium (mg/day)

- Clinical examination
  - Height (cm)
  - Weight (kg)
  - Body Mass Index (BMI)
  - Blue sclerae (Yes/No)
  - Dentinogenesis (Yes/No)
  - Deformity (Yes/No)[1]

- DEXA
  - Location
  - BMD
  - T-Score (Z-Score 18-21 years old only)

- Spine X-ray
  - Vertebral fractures (Yes/No)
    - thoracic and lumbar vertebrae

- The following baseline information will be listed only:
  - Fracture history
  - Bisphosphonate use

[1] If a deformity exists, the location and severity will be listed separately.

No formal statistical testing of the baseline characteristics will be performed.

4.3 Medical History (ITT)

Medical history will be summarised by treatment and overall, presenting number and percentage of patients experiencing each condition (e.g. asthma).

4.4 Adherence with trial protocol and allocated treatment (ITT)

No formal statistical testing will be performed. The following will be presented:

- Numbers of patients who were randomised but never treated.
- A summary of patients attending and withdrawing at each visit will be presented, together with reasons for withdrawal.
- A listing of patients where the blind was broken early will be presented, including details of allocation, timing and reason for breaking blind, and outcome.
- Adherence in the TPTD/ZA group will be summarised using the pen diary information. An estimate of overall % adherence for the duration of the trial will be determined and presented.
- Patients in the standard care arm who are taking bone modifying treatments will be asked for an estimate of compliance each time they have stopped taking the relevant treatment. At the end of the trial, patients will be asked for an estimate of adherence for those medications without an end date. Adherence will be defined as 100% during periods in which standard care patients have not been prescribed bone modifying treatments. Similar to the TPTD/ZA group, estimates of adherence will be collated to provide an overall % adherence to treatment for the duration of the trial. Instances of standard care patients taking TPTD (a prohibited medication) will be recorded.
- TPTD discontinuations will be summarised – type of discontinuation (permanent or temporary), the reason for stopping and the duration of taking TPTD will be presented. Instances of TPTD patients taking prohibited bone modifying treatments will be recorded, as will the adherence to such treatments.

4.5 Concomitant Medication (ITT)

Number and percentage of patients will be presented by medication, treatment and overall, in the categories used in the database. These will be split into prohibited and non-prohibited medications.

4.6 Primary Outcome (ITT)

Statistical analysis of the primary outcome will be based on the ITT population.

The primary outcome is the proportion of participants experiencing a clinical fracture validated by x-ray or other imaging.

Primary analysis

The main analysis of the primary outcome will summarise time to first fracture by treatment group using Kaplan-Meier survival curves.

The proportional hazards assumption will be verified by plotting log-cumulative hazard versus log-time for each treatment group. Model validity will be further explored using plots of Cox-Snell and Martingale residuals.

If the proportional hazards assumption holds, time to first fracture will be analysed using Cox proportional hazards regression. The effect of treatment allocation will be reported as an adjusted (for study site and the minimisation variables[2]) hazard ratio (TPTD vs standard care) with its corresponding 95% confidence interval. Study site will be included as a random effect via a frailty term while the minimisation variables will be included as fixed effects. The unadjusted hazard ratio and corresponding 95% confidence interval will also be presented.

Where the proportional hazards assumption does not hold, treatment groups will be compared using the log-rank test stratified by study site and the minimisation variables[2].

Secondary analysis

A secondary analysis will use binary logistic regression, with treatment group (TPTD vs. standard care), study site (random effect) and the minimisation variables (fixed effects) as the independent variables.

The effect of randomised treatment will be measured by the odds ratio (and 95% confidence interval) for TPTD/ZA vs. standard care.

Sensitivity analysis

Departure from randomised treatment

The primary intention-to-treat approach analyses the participants as randomised i.e. it estimates the effect of the treatment assignment as determined by the randomisation and not by the effect of the treatment that was actually received.

Non-adherence to allocated treatment is likely to exist in both treatment arms in varying degrees. Patients may miss some doses of the treatment or not take any at all for various reasons, some patients will not complete the trial. In order to account for this, a complier-average causal effect (CACE) approach will be taken. We will estimate this with a randomisation-based efficacy estimator, using information on non-adherence and the number of patients with complete information. This approach will avoid selection bias, while allowing the benefit of treatment to vary according to the amount of treatment actually received.

Treatment adherence will be assumed to be binary with a clinically relevant adherence of ≥80% used to define whether a patient is adherent or non-adherent. Baseline propensity to adherence will be estimated based on age, gender and clinical features as recommended in a recent review [2].

[2] The minimisation variables are: clinical fracture during the two years prior to randomisation; OI clinical subtype (type I or others); gender; lowest BMD T score at spine or hip (≤ -2.5 or > -2.5); age (≤ 50 or > 50); bisphosphonate at baseline or within 2 years prior to randomisation.

The primary analysis of time to first fracture will be repeated by adjusting the proportional hazards model for baseline propensity to adherence, as recommended in [3], estimating a causal proportional hazards effect of treatment in the adherent group of patients.

The secondary analysis of the primary outcome using logistic regression will be repeated using the methods recommended in [4]. Patients will be categorised into 'compliance-types' (never-takers; compliers; or always-takers) which can be considered a baseline characteristic and independent of randomised group. The CACE-based odds ratio (and 95% confidence interval) will be estimated using the ITT odds ratio, adjusted for the proportion of non-adherence among TPTD/ZA and standard care group patients who experience a clinical fracture.

Missing data

It is likely that there will be missing data, particularly for the TPTD/ZA group, due to discontinuation of treatment or withdrawals from the study. A sensitivity analysis will be conducted in which missing primary outcome data are imputed. The imputation method will be determined at time of analysis, taking into consideration the choice of estimand and assumptions such as missing at random (MAR) or missing not at random (MNAR) depending on reasons for loss to follow-up. The imputation method will be model-based and shall make use of the minimisation variables as part of the imputation model [5].

4.7 Secondary Outcomes (ITT)

The secondary outcomes for the trial will be reported based on the ITT population as follows:

i. The total number of clinical fractures experienced by participants validated by x-ray or other imaging.

ii. The number of incident vertebral fractures assessed by imaging of the thoracic and lumbar spine.

iii. The total number of fractures experienced by participants defined as the combination validated clinical fractures and vertebral fractures and fractures reported by participants, where imaging was not performed, not feasible or where the results were inconclusive.

Fracture incidence, as defined in i. to iii. above, will be summarised separately by treatment group and overall. Statistical analyses will be by Poisson regression analysis, adjusting for the minimisation variables, and, if required, including an overdispersion parameter.

iv. Bone pain assessed by the brief pain inventory (BPI); quality of life as assessed by the SF36 questionnaire; sleep quality questionnaire (PSQI) and functional status as assessed by the health assessment questionnaire (HAQ) and EuroQol5D (EQ5D) assessment tools, and adverse events.

All quality of life measures will be summarised by treatment group and visit (baseline, 12 months, 24 months, end of study).

With the exception of HAQ, change from baseline in the quality of life measures will be summarised and modelled using an analysis of covariance (ANCOVA) adjusting for baseline

value, study site (as a random effect) and the minimisation variables. The estimated treatment effect and 95% confidence interval will be presented for each outcome.

For the HAQ data, a regression approach will be taken initially. However, since the HAQ data is likely to be heavily skewed with inflation at zero, the chi-squared goodness-of-fit test from the regression model will be assessed and, if statistically significant, then a two-part model will be implemented to account for the anticipated skewed results.

This model will comprise an initial logistic regression to model the probability of a positive score, followed by a second stage modelling the distribution of positive scores using a linear mixed model approach. The model will include the minimisation variables as covariates and results will be presented in terms of treatment effects and 95% confidence intervals.

4.8 Safety Outcomes (Safety population)

The reporting of safety outcomes will be based on the safety population.

i. Adverse Events (AEs) will be summarised by treatment and by severity, causality and seriousness, reporting the both the number of events and the number of patients experiencing a given event. No formal statistical analysis will be performed. A listing will be produced detailing each event, and what happened to the patient subsequently.

ii. Details will be provided of any patients who become pregnant or who have a partner who becomes pregnant during the study. A listing will be produced detailing each event, and what happened to the pregnancies subsequently.

iii. Safety blood results will be summarised by treatment and visit (Baseline, 12 months, 24 months, end of study). Measurements of creatinine (µmol/L), albumin (g/L), calcium (mmol/L), eGFR (ml/min) and alkaline phosphatase (U/L) will be summarised.

iv. Serum 25 (OH) D (nmol/L) will be summarised by treatment and visit (Baseline, 12 months, 24 months, end of study).

4.9 Mechanistic analyses (ITT)

The relationship between each of patient demographics, clinical features of OI, bone density values at baseline, type of genetic mutation and biochemical markers of bone turnover with fracture occurrence and the response to treatment will be explored by means of a subgroup analysis.

The following subgroups are planned:

i. Clinical subtype of OI (type I or others)

ii. Baseline BMD (lowest BMD T score at spine or hip (≤ -2.5 or > -2.5))

iii. Gender

iv. Molecular diagnosis (mutations of COLIA1 or COLIA2 that cause a missense or nonsense [stop] codon in the coding region ["qualitative" mutations]; mutations of COLIA1 or COLIA2 that are

expected to result in a null allele due to nonsense mediated RNA decay ["quantitative" mutations]; mutations in other gene of any type)

Descriptive statistics by treatment group for each subgroup will be presented.

The primary outcome, defined as the proportion of participants experiencing a clinical fracture validated by x-ray or other imaging, will be analysed for these subgroups. The interaction between subgroup and treatment will be included in the primary and secondary analysis models (described in Section 4.5) to determine if the treatment effect differs by subgroup. Grouping of less frequent subgroup categories will be considered where necessary to enable successful model fit.

4.10 Exploratory analyses

The analyses of section 4.9 will be extended to assess the associations among the subgroup variables clinical subtype, baseline BMD, gender and molecular diagnosis. Pairwise associations will be evaluated using cross-tabulations and chi-squared tests.

5. Derived variables

5.1 Brief pain inventory (BPI) [6]

The BPI gives two main scores: a pain severity score and a pain interference score.

The pain severity score is calculated from the four questions (3-6) about pain intensity. Each item is rated from 0 (no pain) to 10 (pain as bad as you can imagine). The scores from the 4 questions are added together and then divided by 4, giving a severity score out of 10.

The pain interference score corresponds to Question 9 responses. The seven sub-items are rated from 0 (does not interfere) to 10 (completely interferes). The scores are added together and divided by 7, giving an interference score out of 10.

Question 2 (pain drawing diagram), Question 7 and Question 8 (pain relief treatment or medication) do not contribute to the scoring.

5.2 SF-36 questionnaire [7]

The SF-36 (v1) Quality of Life questionnaire consists of 36 generic health questions. There are 8 health domains of the questionnaire, each of which are summarised (Physical functioning score (10 items), Role-physical score (4 items), Bodily pain (2 items), General health score (5 items), Vitality score (4 items), Social functioning score (2 items), Role-emotional score (3 items), and Mental health score (5 items)).

The answers to each question (recoded as necessary) are summed for each subject at each visit, within each of the 8 domains. If an item is missing, it will be imputed as the mean of the non-missing items in its domain for the purposes of calculating the domain score.

At least 50% of the item scores in a domain must be non-missing to calculate the domain score, otherwise the domain score is set to missing.

The resulting score for each domain (after the imputation described above) is then standardized, to obtain values ranging from 0 to 100, with higher values indicating a better quality of life.

Two overall summary measures (physical and mental component scores) are then calculated. Further calculation details can be found in [7].

5.3 Pittsburgh Sleep Quality Index (PSQI) [8]

The Pittsburgh Sleep Quality Index contains 19 self-rated questions. The questions are combined to form seven 'component' scores, each of which has a range of 0-3 points.

In all cases, a score of 0 indicates no difficulty while a score of 3 indicates severe difficulty. The seven component scores are added to yield one global score, ranging from 0-21 points, 0 indicating no difficulty and 21 indicating severe difficulties in all areas.

Details of the scoring of the seven components can be found in [8].

5.4 Health assessment questionnaire (HAQ) [9]

The HAQ includes eight blocks of questions (dressing and grooming, arising, eating, walking, hygiene, reach, grip and common daily activities) covering difficulties when performing simple daily activities.

There are 20 questions in total. A 4 point grading system is used to denote degree of difficulty (0=none, 1=some difficulty, 2=great difficulty, 3=not able to perform at all).

In addition, each of the items has a companion aids-devices variable that is used to record what type(s) of assistance, if any, the subject uses for his/her usual activities. These variables are coded from 0 to 3 (0 = No assistance is needed, 1 = A special device is used by the subject in his/her usual activities, 2 = The subject usually needs help from another person, 3 = The subject usually needs BOTH a special device AND help from another person).

The subject must have a score for at least 6 of the 8 categories. If there are less than 6 categories completed, the HAQ score cannot be computed.

The highest score reported for any component question of the eight categories determines the score for that category.

If either devices and/or help from another person are checked for a category and the highest score for this category is 0 or 1 then the score is set to 2. If the basic score is already 2 or 3, the score remains unchanged.

A global score is calculated by summing the scores for each of the categories and dividing by the number of categories answered.

6. Validation and QC

The following will be done by a second statistician:

1. Separate programming and checking of primary outcome results and conclusions.
2. The statistical report will be read and sense-checked.

7. Data sharing

A file, or set of files, containing the final data will be prepared, along with a data dictionary. These will be made available to the Chief Investigator at the end of the analysis phase.

8. References

1. SAS® Institute Inc. SAS for Windows. SAS Institute Inc.: Cary, NC, U.S.A
2. Wheeler KJ, Roberts ME, Neiheisel MB. Medication adherence part two: Predictors of nonadherence and adherence. Journal of the American Association of Nurse Practitioners 2014;26: 225–232.
3. Loeys T, Goetghebeur E, A Causal Proportional Hazards Estimator for the Effect of Treatment Actually Received in a Randomized Trial with All-or-Nothing Compliance. Biometrics 2003; **59**, 100-105.
4. White IR, Uses and limitations of randomization-based efficacy estimators. Statistical Methods in Medical Research 2005; **14** 327-347.
5. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT et al. The prevention and treatment of missing data in clinical trials. N Engl J Med; 367:1355-60.
6. Daut RL, *et al.* Development of the Wisconsin Brief Pain Questionnaire to Assess Pain in Cancer and other Diseases. Pain, 17 (1983), pp. 197-210.
7. Ware JE Jr, Sherbourne CD. "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection,". *Medical Care*, 30:473-483, 1992.
8. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. (1989) The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research 28,* 193-213.
9. Fries JF, Spitz P, Kraines RG, & Holman HR, Measurement of patient Outcomes in arthritis. Arthritis and Rheumatism, 23, 1980, pp.137-145.