Statistical Analysis Plan

TRIAL FULL TITLE	**ANTI-CD3 MAB (TEPLIZUMAB) FOR PREVENTION OF DIABETES IN RELATIVES AT-RISK FOR TYPE 1 DIABETES MELLITUS (Protocol TN-10)**
EUDRACT NUMBER	
Phase	Phase IIb
SAP VERSION	
ISRCTN NUMBER	
SAP VERSION DATE	March 27, 2018
TRIAL STATISTICIAN	Brian Bundy
TRIAL CHIEF INVESTIGATOR	Kevan Herold
SAP AUTHOR	Jeffrey Krischer

Table of Contents

1. Abbreviations and Definitions

AE	Adverse Event
ADA	American Diabetes Association
AUC	Area Under the Curve
BMI	Body Mass Index
DSMB	Data Safety Monitoring Committee
ITT	Intent to Treat
OGTT	Oral Glucose Tolerance Test
PH	Proportional hazards
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
T1DM	Type 1 Diabetes Mellitus

2. Introduction

Preface

The rationale for this study is that individuals with immunologic markers of T1DM and abnormal glucose tolerance are at very high risk for progression to overt disease. They have a condition that differs from overt diabetes only in the duration of the autoimmune process that results in beta cell destruction. It is hypothesized that intervention at the "prediabetic" stage is likely to be more effective than intervention in those in whom frank hyperglycemia has developed and beta cell function has deteriorated further because insulin production is greater before compared to after the diagnosis.

Purpose of the analyses

Analyses of study data will be conducted to address all objectives and other interrelationships among elements of study data of interest to the investigators and of relevance to the objectives of the study. Analyses by sex, age, and race/ethnicity are also planned.
All primary analyses will be conducted under the intention-to-treat principle whereby all outcome data in all randomized subjects will be included, regardless of treatment compliance.

Primary Outcome

The primary outcome is the elapsed time from random treatment assignment to the development of diabetes or time of last contact in the Intention to Treat Population

Criteria for diabetes onset are, as defined by the American Diabetes Association (ADA), based on glucose testing, or the presence of unequivocal hyperglycemia with acute metabolic decompensation (diabetic ketoacidosis)(1,2). For criteria based on glucose testing, one of the following criteria must be met on two occasions as soon as possible but no less than one day apart for diabetes to be defined:

1. Symptoms of diabetes plus casual plasma glucose concentration $\geq$ 200 mg/dL

(11.1 mmol/l). Casual is defined as any time of day without regard to time since last meal. The classic symptoms of diabetes include polyuria, polydipsia, and unexplained weight loss.

2. Fasting plasma glucose ≥ 126 mg/dL (7 mmol/l). Fasting is defined as no caloric intake for at least 8 hours.

3. 2 hour plasma glucose ≥ 200 mg/dL (11.1 mmol/l) on an Oral Glucose Tolerance Test (OGTT). The test should be performed using a glucose load containing the equivalent of 1.75g/kg body weight to a maximum of 75 g anhydrous glucose dissolved in water.

3. Primary Analysis

The study design is a randomized double-blind placebo controlled trial. The primary objective of the TrialNet Anti-CD3 Trial is to assess the effect of teplizumab versus control on the risk of diabetes onset in the intention-to-treat population.

The cumulative incidence of diabetes onset over time since randomization within each treatment group will be estimated using the Kaplan-Meier method (proportion surviving diabetes-free as a function of time). The difference between groups in the cumulative incidence curves, and the associated hazard functions, will be tested at the 0.025 level, one-sided, using the Cox Proportional Hazards (PH) model with discrete time intervals at the 6 month OGTT intervals (3,4). The hazard ratio of diabetes onset between treatment arms will be estimated from the PH model. The critical values will be determined by the group-sequential procedure outlined in the section entitled Interim Monitoring Plan below. The primary test of treatment effect will be adjusted for the design strata and any covariates identified using the procedure outlined below.

Using a step-up procedure additional covariates will be tested and included in the model only if they improve the log-likelihood at 0.10 level (2-sided). This will be accomplished with the treatment assignment variable included but the inclusion/exclusion of the candidate covariates will be completely independent of the treatment variable's impact on the model. The candidate covariates to be tested for inclusion are: sex, BMI, HbA1c, HLA (DR3/4 vs. others), baseline C-peptide (fasting, peak, AUC), baseline OGTT glucose (fasting, 2-hour, AUC), autoantibody presence (mIAA, GADA, IA2A, ZnT8) at study entry. The Wald test associated with treatment variable in the full, adjusted model will be used for the test of treatment effect described in the previous paragraph. Thus, the adjustment of the significance level, as with multiple testing, is unnecessary.

4. Secondary Outcomes and Analyses

The original design of this study anticipated a two-year enrollment period and follow-up of 3 years after the last subject was enrolled for a total of 5 years. To deal with the potential loss of drug effect, the treatment arms will be compared at 5 years, as if the study did follow the original plan (i.e., had the study progressed according to the original plan with the minimum accrual, 50% of the subjects would have been followed for 4 years and 50% for 5 years. For purposes of this secondary outcome, study subjects will be censored at 5 years following their randomization date). This will be the principal pre-specified secondary objective All other planned secondary objective will be considered exploratory.

Exploratory, pre-specified secondary analyses are identified below.

1. For those quantitative baseline factors (including weight, BMI, and the immunologic and metabolic factors, including the autoantibody titers, basal C-peptide, OGTT stimulated C-peptide (peak and AUC mean), and measures of insulin secretion and insulin resistance modeled from the OGTT) entered into the model used in the primary analysis, an attempt will be made to distinguish levels or intervals that correspond to different diabetes risk. The treatment effect within the quantitative levels of each factor will then be assessed through a covariate by treatment group interaction in a PH model. Such an analysis will also be conducted to assess the effects of age as quantitative covariate as described above.

2. Additional exploratory subgroup analyses will be conducted comparing the effects of teplizumab versus control on the risk of diabetes within subsets of the study cohort. Pre-specified subgroups of the enrolled population will be classified by age (8-17 and 18-45) sex, race/ethnicity, autoantibody type and titer at baseline, OGTT in the diabetic range at baseline (yes/no), and HLA risk categories, including the presence or absence of HLA DQB1*0602 and other factors deemed important (such as site and temporal effects) Differences in the treatment effect between subgroups will be tested using a covariate by treatment group interaction term in a PH model (3) and significance levels reported regardless of whether they achieve nominal significance due to the lack of statistical power.

3. Although there is limited power, the diabetes-free proportion at 1,2 and 3 years after randomization will be calculated and compared using the Kaplan-Meier estimates of the proportions and the Greenwood estimates of the variance.

4. Longitudinal analyses will assess the effects of teplizumab versus control treatment on immunologic and metabolic markers over time up to the onset of diabetes. Differences between groups in the mean levels of quantitative factors over time will be assessed using a normal errors linear model for repeated measures. Differences between groups in the prevalence of qualitative factors over time will be assessed using generalized estimating equations for categorical measures. Generalized estimating equations may also be employed for the analysis of quantitative factors when the normal errors assumptions are violated (5).

5. Once a subject develops diabetes, the subject will have reached the primary outcome of the study. However, the subject may still be followed for assessment of other outcomes that will permit continued longitudinal assessment of metabolic and immunologic parameters. This requirement may be satisfied through participation in another study in TrialNet.

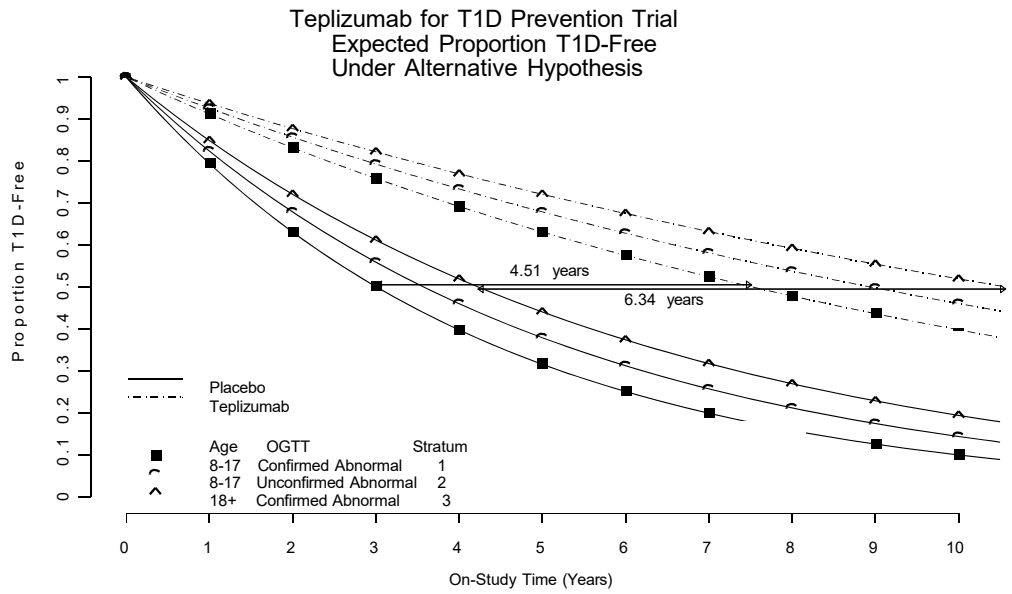
6. The association of demographic, genetic, immunologic, metabolic, and lifestyle factors, among others, both at baseline and over time, with the risk of diabetes onset will be assessed in Cox PH Models over time. The effects of changes in longitudinal factors on diabetes risk will be assessed using time-dependent covariates for these factors. Analyses will be conducted separately within the treatment and control groups, and differences between groups in covariate effects (group by covariate interactions) will be assessed. Assumptions regarding a constant hazard rate will be tested to examine differential efficacy over time.

7. Evidence that the HR is not constant over the period of follow up will be assessed and the risk of diabetes will be assessed in a PH model. The primary test for treatment effect is based on the Cox model, which assumes a proportional hazards between treatment groups. If there is a true treatment effect but it is not proportional over the follow-up period the test will have substantially less statistical power than stated in the statistical section of the protocol. Therefore, to establish a reasonable guide to pursue the possibility of a non-proportional treatment effect we will use the guideline requiring a significance level of 0.10 or less of the Wald test from the standard Cox model. Due to the modest size of this trial, nominal level for significance when testing for violations in proportional hazards will not be specified. Rather, graphical diagnostics using Schoenfeld residuals will be employed to explore evidence for a monotonically decreasing effect of treatment over the follow-up period. Also, plotting the Kaplan-Meier time-to-T1D rates by treatment group on the log-log scale will be used to quantify any diminishing effect of treatment (equal distance separation of curves indicates proportional hazard). We will characterize any such decreasing effect of treatment by model parameterization with monotonically deceasing benefit over the follow-up time. Initially, a treatment interaction with log-transformed time-on-study will be fit to the data and the log likelihood improvement in the fit noted. Other time transforms will be explored only to characterize mathematically the rate of the diminishing effect. Of particular interest will be characterizing the point in follow-up where the hazard ratio is 1. Any variations in proportional hazards other than a monotonically decreasing effect will not be of interest because of the possibility that it is simply random error.

5. Study Power and Sample Size

Applying the eligibility criteria for this study to the data from the Natural History Study (TN-01), hazard and accrual rates were estimated from the TrialNet Natural History/Pathway to Prevention Study (PTP; TN-01) for the three eligible strata: 1) ages 8-17 with a confirmed abnormal OGTT, 2) ages 8-17 with an abnormal OGTT that is not confirmed, (expanded eligibility criteria), and 3) 18 or older with a confirmed abnormal OGTT. Assumptions included a constant risk over time and 25% probability of agreeing to participate. For those subjects that never received a confirmatory OGTT in the PTP study we presumed they are divided by stratum size between stratum 1 and stratum 2 in the same proportions as these groups are in the PTP Study. Likewise, hazard rates for T1DM of these grouped strata were determined by weighted average of the hazard rates from the PTP data. That is, for the 8-17 age groups the 3 strata: a) Abnormal OGTT confirmed: HR=0.1771 and accrual rate: 14.6 (35% by size) b) Abnormal but no confirmatory test: HR=0.3370 and accrual rate: 14.1 (33% by size) c) Abnormal OGTT followed by normal OGTT: HR=0.1222 and accrual rate: 13.5 (32% by size) were weighted to derive the hazard rates for strata 1 and 2. In a similar manner the accrual for 8-17 age groups was determined yielding: 22.0 and 20.3 (multiplied by 0.25, probability of agreeing to participate) for strata 1 and 2, respectively.

The estimated hazard rate is 0.231, 0.194 and 0.164 per year for strata 1, 2, and 3, respectively. The median time to T1DM onset for the control group based on a constant hazard rate is 3.01, 3.57, and 4.23 years for strata 1, 2, and 3, respectively. The effect size for this trial is a 60% reduction in the risk of T1DM (i.e., hazard ratio of experimental to control equals 0.4). This reduction in risk translates into a median time to T1DM of 7.52, 8.93, and 10.6 years for the teplizumab group for strata 1, 2, and 3, respectively (the increase is 4.51, 5.36, and 6.34 years, respectively). These design characteristics are displayed in the graph below.



The primary hypothesis test will be the Wald test of the treatment assignment variable when modeling the time to T1DM using the Cox model adjusting for baseline age and OGTT status (expanded eligibility criteria). To achieve statistical power of 80% for a one-sided Wald test at the 0.025 significance level and the effect size described above, will require enrollment and follow-up of enough participants to observe 40 subjects with T1DM onset (38) (this is the "event" sample size in contrast to the study sample size). This event sample size reflects the combination of the study sample size and the amount of follow-up at which the fixed-sample primary hypothesis test may be conducted. Although group sequential testing will be employed, the method of Lan and DeMets maintains the power while controlling the type I error used in determining the fixed sample size.

Participants <18yo will undergo an OGTT prior to the first infusion. The results of this OGTT will be incorporated into the analysis of the primary outcome variable but will not be used to determine eligibility for the study. The study sample size and duration are variable when fixing the "event" sample size. In the absence of any safety concerns and evoking any stopping rules, closing accrual should not occur until sufficient participants are accrued so that projections (based on the observed T1DM rates and the actual accrual pattern) indicate that within a reasonable follow-up period the event sample size will be achieved. The constant hazard rate assumption is retained to compute the initial projection. The projected annual accrual is 5.5, 5.1 and 1.6 for strata 1, 2 and 3, respectively. Allowing for a 5% per year drop-out rate and the approximately 3 dozen subjects already enrolled, the study will need to accrue a total of 71 subjects in 3 years and follow all those enrolled for another 4 years beyond the last enrolled subject to achieve a statistical power of 80% (39).

Note the accrual period and the study sample size are only projections since the actual accrual rate, the control hazard rate and the loss to follow-up rate are estimates from the PTP Study or other similar trials. Furthermore, the over-all hazard rate is sensitive to the age distribution of the enrolled study population, which is also an approximation. As the study progresses, projections of the study duration will be computed based on the observed data (noncomparative treatment analysis) and if in conflict, will be brought to the attention of the DSMB and the TrialNet governing body to determine the best course of action.

6. Interim Monitoring Plan

Interim analyses will be conducted periodically during the study and will be reviewed by the TrialNet Data and Safety Monitoring Board (DSMB) for assessment of effectiveness and safety; the TrialNet DSMB meets at least every six months to review study progress and safety. An independent medical monitor will closely monitor the events in the trial as described in section 10.4. If a group sequential stopping boundary is crossed, the DSMB may terminate enrollment into the trial early. The Lan-DeMets (7) spending function with an O'Brien-Fleming boundary will be used to protect the type I error probability for the primary outcome analyses, and to assess the significance of the interim results periodically during the trial. The spending function that approximates the O'Brien-Fleming boundaries is:

$$\alpha_1(t^*) = 2 - 2\Phi\left[\frac{Z_{\alpha/2}}{\sqrt{t^*}}\right]$$

where t^* is the information fraction ($0 < t^* \leq 1$) and α is the fixed-sample type I error (i.e., 0.025).

The DSMB will also be informed if there is a serious lack of evidence of a treatment effect (i.e. futility analysis). The boundaries are based on the paper by Lachin (41). The study should be stopped based on the futility of rejecting the null hypothesis at the completion of the trial if: $Z_{HR}(t^*) \geq 0$ when $0.5 \leq t^* < 0.8$ or if $Z_{HR}(t^*) \geq -0.8$ when $t^* \geq 0.8$.

Using Lachin's formulas a onetime use of either boundary for the design parameters above ($\theta \equiv Z_{1-\alpha} + Z_{1-\beta} = 2.8$) raises the type II error to approximately 0.204 and 0.202, respectively. For larger values of t^* the increase to the error probability is even less. Furthermore, by the laws of

probability a single use of each rule will increase the type II error no more than the sum of the increase (i.e., 0.15+0.004+0.002 = 0.156).

Additional analysis will assess potential adverse outcomes of treatment and will assess the incidence of all severe adverse events.

7. General Considerations

1. Analysis Populations

The Intention to Treat Population (ITT)

The intention to treat population comprises all randomized (as planned) subjects.

Full Analysis Population

The Full Analysis Set (FAS) will comprise all subjects who received any study drug and who participated in at least one post-baseline assessment. These will be analyzed as randomized. FAS will be the primary efficacy population. So, FAS is a subset of ITT.

Per Protocol Population

The Per Protocol Set (PPS) will comprise all subjects who did not substantially deviate from the protocol as to be determined on a per-subject basis before data base lock and unblinding.

Safety Population

All subjects who received any study treatment (including control) but excluding subjects who drop out prior to receiving any treatment.

During accrual, 2 subjects enrolled on the >= 18 year of age stratum did not have a second OGTT prior to randomization to confirm abnormal glucose tolerance. Eligibility is technically unknown without the results of the missing test. Given that the these subjects had an OGTT on study indicated abnormal glucose tolerance, not necessarily consecutive, we considered these conditions sufficient to retain these subjects for analysis.

2. Timing of Analyses

The final analysis will come after sufficient numbers of events (40 or more) have been reported in the FAS population to achieve the planned 80% statistical power for the primary analysis.

3. Adjustment of Confidence Intervals and p-values

There was 1 interim analysis conducted when 18 subjects were diagnosed with T1D. We applied the Lan-DeMets decision rule as outlined in the protocol and spent type I error of p = 0.00083, one-sided. Therefore, the final analysis should be conducted at 0.0242 level (to preserve type I error). Given this almost negligible difference from 0.025, all p-values will be reported as if a fixed-sample size test was conducted and only noting the adjusted critical value p-value in the results section if the primary hypothesis test falls within the narrow range (i.e., 0.0242 – 0.025).

4. Missing Data

In general, missing values will be assumed to be *missing completely at random (MCAR)* unless empirical evidence to the contrary can be established internal to the study. The methodology employed in analyzing time-to-T1D utilizes whatever follow-up has been recorded for each subject (i.e., maximum utilization of follow-up time). Presuming no evidence against MCAR, and the modest size of the trial, no methods will be employed to impute additional follow-up of subjects that drop out (i.e., lost to follow-up). All secondary endpoints will use the complete-case analysis approach which limits the analytical cohort to those subjects that have the secondary endpoint of interest measured and recorded. In modeling to adjust for risk factors associated with the endpoint (i.e. covariates), missing values will be assigned the mean from the known covariate cohort. This simple rule will be employed only if the percent missing is less than 10% for the analytical cohort. If the missing is 10% to 20% a separate indicator for missing will be included in the modeling. If the missing in exceeds 20% the covariate will be removed from consideration.

8. Safety Analyses

Safety will be evaluated with summary of adverse events for the safety population. The following parameters will be assessed during the study:

Adverse Events
The summary statistics will be produced in accordance with Section 8.
Treatment emergent adverse events (AEs) are those events that occur after the baseline assessment. Only Grade 2 or greater adverse events were reported in this study. The incidence of the following AEs will be reported:

A tabular summary of AE will present: Number of subjects with any AE; Number of SAEs with outcome death; Number subjects with SAE; Number subjects with AEs leading to discontinuation of study drug, even if by protocol; Number of subjects with AEs leading to discontinuation of study; Total number of AEs; Total number of SAEs [TABLE].

The Adverse Events summary tables will include number of adverse events, the number of subjects in each treatment group in whom the event occurred, and the incidence of occurrence and should be grouped by system organ class, preferred terms and/or other interested variables (e.g., relatedness, intensity and seriousness). [TABLE]

When calculating the incidence of adverse events, or any sub-classification thereof by treatment, time period, severity, etc., each subject will only be counted once and any repetitions of adverse events will be ignored; the denominator will be the total population size.
Deaths, Serious Adverse Events and other Significant Adverse Events

All formal testing of adverse effects will be based on the subject as the experimental unit. Thus for comparing incidence of AE within system organ by treatment group, a one-sided Fisher's exact test will be conducted at 0.05 level (higher incidence in experimentally treated group is the alternative hypothesis). Also, highest AE grade will be determined for each subject and compared by treatment group using a 2 sample Wilcoxon test (one-sided at 0.05). No correction for multiple testing will be employed in order that the statistical power is maintained.

9. Reporting Conventions

P-values ≥0.01 will be reported to 2 decimal places; p-values less than 0.01 and >0.001 will be reported to 3 decimal places; p-values less than 0.001 will be reported as "<0.001". The mean, standard deviation, and any other statistics other than quantiles, will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data. Estimated parameters, not on the same scale as raw observations (e.g. regression coefficients) will be reported to 3 significant figures.

10. Per Protocol Analysis

Quantifying the evidence of any dose response relationship is part of a complete analysis of any well run and completed clinical trial. This is especially true when the trial's primary outcome is negative to explore whether there is evidence that deviations from the treatment protocol may have played a role in the negative outcome. Given this we plan to assess the treatment hazard ratio by the degree of compliance to the protocol scheduled dose in a quantitative manner.

Using the Cox model we will assess the evidence for an effect of treatment compliance including the entire cohort. The number of infusions of treatment and/or the treatment dose will be introduced into the model to determine its effect on risk. If there is evidence that it is predictive (<=0.05, one-sided) then a treatment-infusions of therapy, or dose, interaction term will be introduced to see if there is a different compliance gradient between the two treatment groups. The procedure for including covariates, such as age, will follow the set up procedure as described above under Primary and Secondary analyses.

11. Technical Details

The analysis will be performed in R, S-Plus or SAS.

The distributional assumptions as well as other assumptions underpinning the planned analyses will be checked. Final decisions regarding analysis methods and choice of explanatory variables will be taken then.

References

1. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. National Diabetes Data Group. *Diabetes* 28:1039-1057, 1979

2. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* 26:3160-3167, 2003

3. Cox DR: Regression model and Life Tables *J R Stat Soc* 34B:187-220, 1972

4. Kalbfleisch JD, Prentice RL: The statistical analysis of failure time data., 1980

5. Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL: *Analysis of longitudinal data*. Oxford, UK, Oxford University Press, 2002

6. Reboussin DM, DeMets DL, Kim KM, Lan KK: Computations for group sequential boundaries using the Lan-DeMets spending function method. *Control Clin Trials* 21:190-207, 2000

7. DeMets DL, Lan G: The alpha spending function approach to interim data analyses. *Cancer Treat Res* 75:1-27, 1995

8. Wieand S, Schroeder G, O'Fallon JR: Stopping when the experimental regimen does not appear to help. *Stat Med* 13:1453-1458, 1994