Statistical Analysis Plan

TRIAL FULL TITLE	**Anti-Thymocyte Globulin (Thymoglobulin®) and Pegylated GCSF (Neulasta®) in New Onset Type 1 Diabetes (Protocol TN-19)**
EUDRACT NUMBER	
Phase	Phase IIb
SAP VERSION	
ISRCTN NUMBER	
SAP VERSION DATE	August 30, 2017
TRIAL STATISTICIAN	Brian Bundy
TRIAL CHIEF INVESTIGATOR	Michael Haller
SAP AUTHOR	Jeffrey Krischer

Table of Contents

1. Abbreviations and Definitions

AE	Adverse Event
ADA	American Diabetes Association
ANCOVA	Analysis of Covariance
AUC	Area Under the Curve
BMI	Body Mass Index
DSMB	Data Safety Monitoring Committee
ITT	Intent to Treat
MMTT	Mixed Meal Tolerance Test
PH	Proportional hazards
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
T1D	Type 1 Diabetes

2. Introduction

2.1 Preface

A phase 2 study in humans with new-onset T1D and funded by the Immune Tolerance Network (ITN) was completed to determine if ATG (6.5mg/kg) could preserve C-peptide in new onset T1D. The study tested the hypothesis that selective deletion of lymphocytes would reset the immunologic rheostat, effect dynamic immune regulation and perhaps induce and maintain tolerance in T1D. While this study helped to establish the relative safety of ATG in humans with T1D, the study failed to demonstrate benefit (1). Post-hoc analyses have suggested that an initial decline in beta cell function amongst those who were treated with active drug was followed by a relative preservation of beta cell function. We hypothesize that the initial decline in beta cell function was related to the severity of cytokine release syndrome and serum sickness. Furthermore, we hypothesize that the currently proposed protocol (utilizing ~1/3 of the ATG dose in combination with GCSF) will promote a more tolergenic state while inducing less severe cytokine release/serum sickness and, as is supported by our preliminary data, preservation/improvement in beta cell function 12 months after therapy.

2.2 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 and race/ethnicity, as appropriate, 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.

2.3 Primary Outcome

The primary outcome of each participant is the area under the stimulated C-peptide curve

(Y_{AUC}) over the first 2 hours of a mixed meal glucose tolerance test conducted at the one-year visit. The Y_{AUC} is computed using the trapezoidal rule which translates into a weighted sum of the timed C-peptide values over the 120 minute MMTT. The more appealing quantity is the AUC mean which is $\frac{Y_{AUC}}{120}$ in nmol/L and will be denoted as Y_{Cp}.

3. Primary analysis

Let $Y_{Cp}^{ATG/GCSF}, Y_{Cp}^{ATG}$, and $Y_{Cp}^{Control}$ represent the C-peptide AUC mean for study patients receiving both ATG and GCSF, receiving ATG alone, and those receiving placebo, respectively. Likewise, let $\mu_{Cp}^{ATG/GCSF}, \mu_{Cp}^{ATG}$, and $\mu_{Cp}^{Control}$ represent the population mean of C-Peptide for these groups at one year, respectively.

The primary statistical hypotheses to be assessed in the study are:

$$\mathrm{H}_0\colon \mu_{Cp}^{ATG/GCSF} = \mu_{Cp}^{Control} \text{ versus } \mathrm{H}_a\colon \mu_{Cp}^{ATG/GCSF} > \mu_{Cp}^{Control}$$

$$\mathrm{H}_0\colon \mu_{Cp}^{ATG} = \mu_{Cp}^{Control} \text{ versus } \mathrm{H}_a\colon \mu_{Cp}^{ATG} > \mu_{Cp}^{Control}$$

The primary analysis will be conducted on a transformed scale using the function $\log(Y_{Cp}+1)$. This provides better normal distributional behavior by the test statistic. The comparison between any two treatment arms will be based on a Wald test of treatment effect in an ANCOVA model adjusting for gender, baseline age, and baseline (from the adjusted linear model) at the 0.025 level (one-sided). Thus the overall Type I error for both tests is approximately 0.05.

Using a step-up procedure additional covariates will be tested and included in the model only if they improve the log-likelihood at 0.05 level (2-sided). This will be accomplished with the treatment assignment variable included but the inclusion/exclusion of the candidate covariate will be completely independent of the treatment variable's impact on the model. The Wald test associated with treatment variable in the full adjusted model will be used only as confirmation of the hypothesis test of treatment described in the previous paragraph. Thus the adjustment of the significance level is superfluous.

4. Secondary Outcomes and Analyses

Additional analyses of the primary outcome to determine the effect of ATG with and without GCSF include:

- C-peptide AUC mean at 3, 6, 9, 12, 18 and 24 months using the ANCOVA mentioned above.

- The hazard rate of C-peptide failure (C-peptide failure defined as the first occurrence at which the 2 hour peak C-peptide < 0.2 nmol/L during a MMTT) using the proportional hazards model while adjusting for baseline level of C-peptide, gender and baseline age.

- Longitudinal analyses of C-Peptide AUC means 3, 6, 9, 12, 18 and 24 months using a mixed effects model with a random intercept and slope by subject, adjusted for the baseline level of C-peptide, sex and baseline age. The mean intercept and slope will be compared between treatment groups.

- Treatment interactions with the covariates baseline C-peptide, sex and baseline age will be analyzed with a homogeneity test categorizing the continuous variables into 3 approximately equal groups [could substitute age groupings recommended by investigators]; ladder plots will be constructed. Other variables to be tested for treatment interactions are HbA1c levels, HLA, other genotype and immune phenotypes, and race/ethnicity, as appropriate.

Additional secondary objectives to determine the effect ATG with and without GCSF has on the following:

- HbA1c, Insulin dose (units/kg) and blood glucose over time by treatment group using ANCOVA.
- Adverse events
    - Number and severity.
    - The rates of severe adverse events will be computed (total number of events divided by total subject years of follow-up).
- Hypoglycemia
    - Number of major hypoglycemic events (defined as loss of consciousness, seizure, or requiring assistance from another person because of altered state of consciousness).
    - Reported hypoglycemic events confirmed with capillary blood glucose measurement less than 70 mg/dl.
    - The rates of severe hypoglycemic will be computed (total number of events divided by total subject years of follow-up).

Additional Outcomes and Analyses

A mechanistic analysis plan will explore the following hypotheses:

- Effector T cell phenotypes are depleted upon treatment
- Treg phenotypes are partially spared upon treatment
- Hypo-responsive T cell phenotypes are increased upon treatment
- Tolerant/regulatory APC phenotypes correlate with kinetic changes in tolerance T-cell phenotypes
- Favorable changes in one or a combination of the above may correlate with treatments success (clinical outcome TBD).

To evaluate these hypotheses, the ITN will perform an extensive phenotypic characterization of T and APC cells using already developed flow panels to evaluate the impact of ATG+GCSF v ATG treatments on their relative frequencies and their activation status in this trial. These same panels were previously tested in T1DAL and START (T cell panel) and AbATE (T cell and APC panels) trials; ITN will also be able to compare T cell and APC responses across therapeutic strategies in these T1D trials.

Flow Panels:

Modified X-trial T cell panel	APC panel
CD56	CD19
Granzyme B	CD274 (PD-L1)
L/D	CD141 (BDCA-3)
CD8	CD3
CD57	CD304 (BDCA-4)
Eomes	HLA-DR
CD3	CD123
FOXP3	CD11c
CD279 (PD-1)	CD1c
TIGIT	CD16
CD197 (CCR7)	CD4
CD127 (IL-7Ra)	CD14
CD45RA	CD86
KLRG1	CD56
CD45RO	
CD4	

A minimum of 5 million cryopreserved PBMC will be provided from randomized study participants at study weeks 0, 2, 10, and months 6 and 12 for longitudinal batched assessment.

Based on the clinical outcome of the trial and/or successful identification if immune biomarkers of treatment and/or response outcome from this initial flow cytometric analysis, future immunophenotyping experiments with additional study time points may be proposed to further characterize cell populations of interest, an explore the kinetics of immune signatures of treatment and/or clinical response.

Given the unknown expectations and interpretive power of the proposed analyses, the study team understands that the results may be somewhat limited; however, early positive findings would reinforce the primary publication so the team agreed that this the most practical way to move forward.

Additional analyses will compare the results in this trial to other trials using ATG and GCSF and other TrialNet studies. Data in this trial will be used in conjunction with other TrialNet data for exploratory analysis.

5. Sample Size and Power Calculations

The primary analysis will compare the difference in C-Peptide between experimental and placebo treatment groups at 12 months using $\log(Y_{Cp}+1)$ transformation and ANCOVA model adjusting for sex, baseline age and the baseline value of $\log(Y_{Cp}+1)$. Estimates of the mean and standard deviation of $\log(Y_{Cp}+1)$ (expressed algebraically as: $\hat{\mu}_{\log(Y_{Cp}+1)}$ and $\hat{\sigma}_{\log(Y_{Cp}+1)}$) in the placebo group were derived from the last four TrialNet studies of early onset disease. The 90% confidence bound estimates were used to provide good confidence that this trial's study population characteristics fall within these limits to assure the advertised statistical power; the estimates are $\hat{\mu}_{\log(Y_{Cp}+1)}=0.360$ and $\hat{\sigma}_{\log(Y_{Cp}+1)|X}=0.167$ (i.e., the residual mean squared error regressing on sex, baseline age and C-peptide). The geometric-like mean of Y_{Cp} for the placebo group is $\exp(0.341)-1=0.406$ nmol/L.

Using standard equations for the comparison of two means and a 1:1:1 allocation, a sample size of 78 participants (26 per group) with complete data would provide power of 85% to detect a 50% increase in the geometric-like mean in either experimental treatment group (compared to the control) using the Wald test (from the adjusted linear model) at the 0.025 level (one-sided). Thus the overall Type I error for both tests is approximately 0.05.

Assuming that 10% of the participants will have missing data (one-year MMTT was not done or subject withdrew prior to the one-year assessment), the sample size goal for this study will be set at 84 participants (28 per group).

The study will be closed to additional participants when the total number then randomized plus a fraction of those in screening (i.e., screening for eligibility) is expected to provide the proper number of eligible participants. Participants who had already conducted the initial screening visit at that time will be allowed to complete screening and be randomized if both consenting and eligible.

In the situation where both hypotheses are rejected in favor of the experimental regimen, then it is appropriate to decide which experimental regimen should be considered first for additional clinical trials. If ATG and ATG+GCSF truly differ in efficacy, it is not expected that the difference would be as large as the effect size used in this design. Thus requiring the difference to reach statistical significance would be too stringent due to the lack of statistical power. The plan is to employ the method of Simon[2] to select the treatment with the largest geometric-like mean in C-Peptide regardless of how small the difference is over the other experimental regimen. Formally, the decision rule is to select the experimental regimen with highest predicted mean based on the fitted linear regression model of C-Peptide used in the formal hypothesis tests. Given the one experimental treatment is associated with 25% higher mean (half of the design effect size) compared to the other experimental treatment, the probability is 96% that the most efficacious regimen will be selected with this decision rule. Even if the increase is only 12.5% (i.e., a fourth of the design effect size), the probability is 82% in making the right choice. Since under the null

($\mu_{Cp}^{ATG} = \mu_{Cp}^{ATG/GCSF}$) we consider selecting either treatment a non-error there is no contribution to the type I error, and therefore, no further adjustment is required to the α-level of the initial two pair-wise tests.

Sample Size Re-estimation

The Residual Mean Square Error (RMSE) and the mean of the control group are two parameters used in determining the sample size goal (N_{SSG}=78) which are directly tied to the study cohort enrolled. The former value is directly proportional to the sample size and the latter applies its influence in defining the effect size. Thus the advertised statistical power of this trial may differ from the actual power if either differs from the initial values assumed. Consequently, the plan is to estimate both values using the accumulated data (internal interim estimate, IIE) when approximately half the subjects (N_{SSG}/2) have had their 1 year C-Peptide assessment. Since the initial values used are based on real data from four previous TrialNet studies, the re-estimation of each parameter will be a two-term weighted average using the number of subjects on which the IIE is based as the one weight (N_{IIE}) and the remaining number of subjects participants to be enrolled as the other weight (N_{SSG}-N_{IIE}). Algebraically, the re-estimate will be computed:

$$\hat{\sigma}^2_{Re-estimate} = \frac{N_{IIE}\hat{\sigma}^2_{IIE} + (N_{SSG} - N_{IIE})\hat{\sigma}^2_{Initial}}{N_{SSG}}$$

Similarly, re-estimation of the mean of the control group will be calculated. These new estimates will be used to calculate a new sample size goal for the trial.

It is important to note that this adaptive procedure is a non-comparative interim analysis (i.e., the observed treatment effect has no influence on the sample size re-estimation). Therefore, this analysis has no effect on the type I error. Note the re-estimation can be conducted by the analyst in a blinded fashion. The goal is to assure adequate statistical power at the completion of the trial by assuring the two design parameters reflect accurately the study population being enrolled.

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[3] 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 monitoring plan will allow for early termination based on the treatment effect on C-peptide values at 1 year of follow-up using the ANCOVA model described above.

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[4]. The study arm should be "closed" based on the futility of rejecting the null hypothesis at the completion of the trial if:

$z_{ATG}(t^*) \leq 0.1$ when $0.5 \leq t^* < 0.75$ or if

$z_{ATG}(t^*) \leq 1.0$ when $t^* \geq 0.75$ (z is the Wald test of the treatment effect coefficient). The same rule would be applied to the ATG/GCSF treatment group. These $t^* = 0.5$ and 0.75 are equivalent to when there are 39 and 59 participants with one-year C-peptide results, respectively. Lachin showed that a onetime use of either boundary for the design parameters above ($\theta \equiv Z_{1-\alpha} + Z_{1-\beta} = 3.00$) raises the type II error to approximately 0.15414 and 0.15611, 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 re than the sum of the increase (i.e., 0.00414+0.00611 = 0.0103).

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.

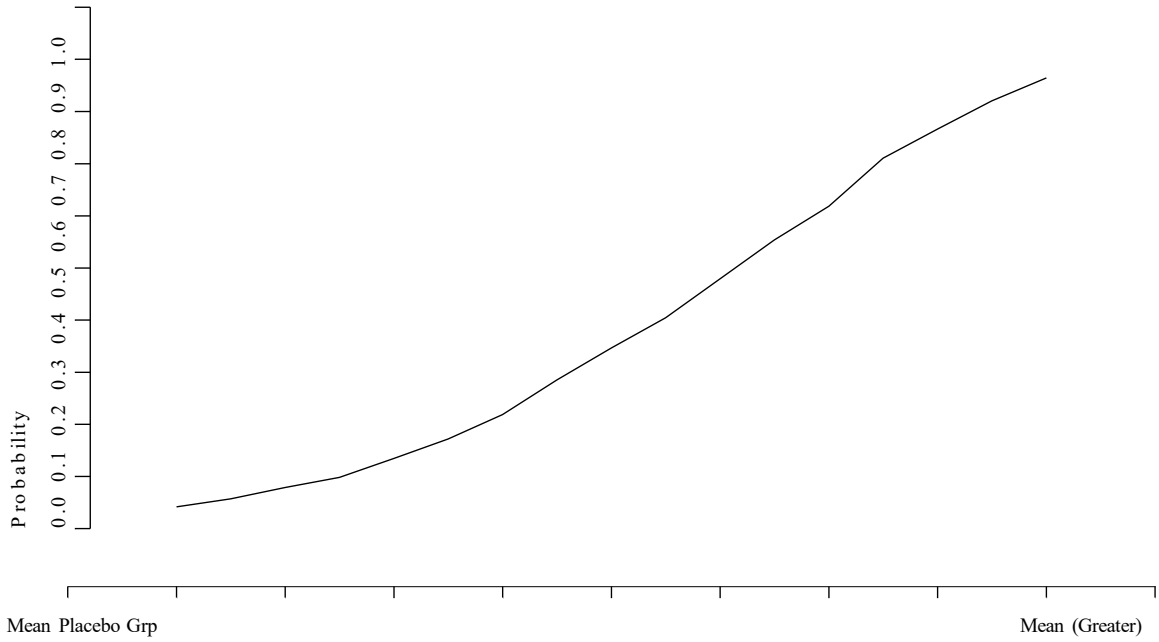
Interim Analysis Study Modification

Whether Lan-DeMets boundary is crossed or a futility decision rule indicates closure of an experimental arm, the course of action for the study is as follows. In case the futility decision rule indicates closure of one of the experimental arms (and the governing bodies concur), the trial should continue randomizing between the other remaining experimental treatment and placebo (1:1 ratio) since the corresponding hypothesis remains unaddressed. However, in the circumstance where the Lan-DeMets boundary is crossed the normal course of action would be to close the placebo arm. This would incapacitate the second formal hypothesis test thus leaving only the issue of which experimental treatment is "best" as the remaining question. However, it may or may not be reasonable to make this decision as well. To evaluate the evidence at the interim the plan is to use stochastic curtailment and compute the probability that the currently 'best' treatment (as defined by the decision rule above) is not the 'best' when the full sample size is reached. In the normal employment of stochastic curtailment in futility

analysis the parameters used for the simulation are based on the alternative hypothesis. The analogous parameters for this situation is to set the mean of the currently ‘best’ treatment to be lower than the other experimental treatment mean and determine the probability of a reversal. As with the normal application of stochastic curtailment, we will use the probability of 0.20 or less to stop accrual completely and invoke the decision rule for selecting the ‘best’ treatment immediately. Alternatively, the study would continue to its planned completion, randomizing between the two active treatment arms, until the projected probability of a re-ordering of the outcome (change in the selection of the superior treatment arm) is less than 0.2.

Our choice for setting the parameters for simulation is to set the ‘best’ treatment group at 25% lower than the other experimental treatment mean; these two parameters being centered at the midpoint between the two estimated experimental group means. This is consistent with the difference between the two treatment arms specified in the Simon selection at the completion of the trial (described above), should both arms be superior to the placebo.
The graph below provides an estimated probability of a ‘best’ selection reversal conditioned on the interim data (simulated under the condition of crossing the Lan-DeMets boundary and t* = 0.5). The probability is primarily a function of the relative position of the ‘lesser’ mean to the positions of the placebo mean and the ‘greater’ mean (all means being estimates from the interim data). That is, the closer the ‘lesser’ mean is to the placebo, the less likely there could be a reversal in the selection of the ‘best’ treatment.

Probability of Reversal Given the Relative Location of the Lesser Experimental Group Mean



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

7. General Considerations

7.1 Timing of Analyses

The final analysis will come after all subjects have had their 1-year MMTT (or been followed long enough to determine they will not) to achieve the planned 80% statistical power for the primary analysis.

7.2 Analysis Populations

7.2.1 The Intention to Treat Population (ITT)

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

7.2.2 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

7.2.3 Complete Case Population

The complete case analysis will comprise all subjects that have had the endpoint measured. This is the population on which the primary hypothesis test will be based.

7.2.3 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. These will be analyzed according to actual treatment received and stratum.

7.2.4 Safety Population

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

During accrual we discovered that a couple of 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 the slow accrual of this trial we pursued every reasonable justification to retain such subjects blinded from their end result. 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. This rule will be applied consistently to any subjects >=18 that fail to have the confirmatory test.

7.2.5 Adjustment of Confidence Intervals and p-values

No adjustments will be necessary because no interim analyses were conducted.

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:

8.2 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 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; 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.

8.3 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 ≥0.001 but less than 0.01 will be reported to 1 decimal place; 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 (planned or not) may have played a role in the negative outcome. Given this we plan to assess the treatment effect on the c-peptide mean by the degree of compliance to the ideal scheduled dose in a quantitative manner. Using the ANCOVA model we will assess the evidence for an effect of treatment compliance including the entire cohort. The number of courses of treatment will be introduced into the model to determine its effect on risk (regardless of treatment group). If there is evidence that it is predictive (<=0.05, one-sided) then a treatment-courses of therapy 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.

12. References

1. Gitelman SE, Gottlieb PA, Rigby MR, Willi SM, Fisher LK, Moran A, Gottschalk M, Moore WV, Pinckney A, Keyes-Elstein L, Aggarwal S, Phippard D, Sayre PH, Ding L, Bluestone JA, Ehlers MR, Felner EI, the START Study Team. Antithymocyte globulin treatment for patients with recent-onset type 1 diabetes: 12-month results of a randomised, placebo-controlled, phase 2 trial. *The Lancet Diabetes & Endocrinology*, 28 August 2013; doi:10.1016/S2213-8587(13)70065-2.

2. Simon, R., Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treatment Reports 1985; 69: 1375-81.

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

4. Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*, 24, 2747-2764, 2005.

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

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

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

8. 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

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