# Statistical Analysis Plan

## Study M16-100

## THE EFFICACY OF TESTOSTERONE REPLACEMENT THERAPY IN MIDDLE-AGED AND OLDER HYPOGONADAL MEN WITH MID-LIFE-ONSET, LOW-GRADE PERSISTENT DEPRESSIVE DISORDER (DYSTHYMIA)

**Date: 01 January 2023**

**Version 4.2**

## List of Abbreviations

| | |
|---|---|
| CV | Cardiovascular |
| ESC | Executive Steering Committee |
| EAS | Efficacy Analysis Set |
| FAS | Full Analysis Set |
| GDS-15 | 15 Item Geriatric Depression Scale |
| HIS-Q | Hypogonadism Impact of Symptoms Questionnaire |
| MACE | Major Adverse Cardiac Event |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MICE | Multiple Imputation by Chained Equations |
| PHQ-9 | Patient Health Questionnaire |
| PDD | Persistent Depressive Disorder |
| SAP | Statistical analysis plan |
| TRT | Testosterone Replacement Therapy |

# 1.0 Introduction

This document provides the detailed Statistical Analysis Plan for a study to determine the Efficacy of Testosterone Replacement Therapy in Middle-aged and Older Hypogonadal Men with mid-life onset persistent depressive disorder (**PDD**), pure dysthymic syndrome (hereafter **mid-life onset dysthymia**). This TRAVERSE-PDD study is a substudy of the Testosterone Replacement therapy for Assessment of long-term Vascular Events and efficacy ResponSE in hypogonadal men (**TRAVERSE**) Trial.

# 2.0 Study Background

Multiple lines of evidence support the hypothesis that a major subset of persistent depressive disorder – defined here as mid-life onset dysthymia - might be responsive to testosterone replacement in men.[1] First, epidemiological and clinical studies have demonstrated an association between lower testosterone levels and chronic low-grade depression.[2;3] Second, development of mid-life-onset depression is associated with low testosterone levels in men,[4;5] and in one open-label study, testosterone was an effective antidepressant in late-onset but not early-onset depression.[6] Third, a growing body of evidence supports the possibility that the age-related decline in male testosterone level can be associated with a syndrome that has significant symptom overlap with low-grade depression,[1;7;8] and that related depressive symptoms (e.g., dysphoria, fatigue, loss of vigor, and low libido) improve with testosterone replacement.[7;9-18] Finally, in clinical studies using exogenous testosterone in depressive disorders, the cumulative evidence suggests that testosterone, whether administered alone or as an augmentation strategy in conjunction with traditional anti-depressant therapy, tends to be ineffective in men with major depressive disorder (**MDD**).[19;20 21;22] However preliminary evidence suggests that androgenic therapy can improve depressive symptoms in men with low-grade depression.[23-25]

These data, coupled with our clinical experience administering exogenous testosterone to depressed men, suggested to us that testosterone replacement might be efficacious as a treatment for this subgroup of men with PDD, pure dysthymic syndrome (previously referred to as dysthymia in DSM-4), but not in men with MDD. Specifically, we propose that since mid-life-onset, dysthymic PDD appears to be a relatively common, distinct clinical entity in hypogonadal men, and is likely to be particularly unresponsive to established antidepressant interventions[26;27], the clinical investigation of testosterone replacement as a specific and novel antidepressant therapy in such men is warranted.

The taxonomy, epidemiology, and clinical course of chronic depression remains poorly understood. Prior to DSM-5, those with low-grade chronic depression were categorized under the heading of "dysthymia." This term has been replaced in DSM-5 by the broader term PDD.[28;29] This new construct in DSM-5 includes patients exhibiting at least two years of relatively persistent depressive symptoms, regardless of whether these symptoms are severe (i.e., meeting full criteria for a major depressive episode: low mood or loss of interest all day every day, with associated neurovegetative symptoms), or moderate (i.e., low mood most of the day, more days than not, with some associated neurovegetative symptoms but not of sufficient severity to meet criteria for a major depressive episode).

PDD thus represents a heterogeneous category comprised of some patients with chronic major depressive disorder (MDD) and others with a milder chronic depressive disorder, termed PDD with pure dysthymic syndrome. The age of onset is apparently important in distinguishing sub-groups of chronic depression. Specifically, those with an early onset of chronic depression have an MDD-like condition (whether or not current symptoms are low-grade or severe); whereas those with a middle-age-onset of chronic, low-grade depression have a fundamentally different condition. Further, this latter type of chronic depression is often associated with an age-related, progressive medical condition, such as cerebrovascular disease[30], a dementing illness[31], type 2 diabetes mellitus[32], or hypogonadism.[2;33;34] We hypothesize that a distinct subgroup of PDD: low-grade, mid-life-onset male depression is associated with hypogonadism and responsive to testosterone therapy. For the purposes of this document, **we use PDD to refer specifically to this mid-life onset category of dysthymic-type persistent depressive disorder**, of which TRAVERSE participants are at risk.

Accumulated evidence from studies of dysthymia (i.e., low-grade, chronic depression) supports this distinction. Patients with an early-onset of dysthymia overlap in virtually all measures with MDD: they typically have a positive family loading for depressive illness, have periods of time when they have met full criteria for MDD, exhibit the usual treatment response to antidepressant interventions, and importantly, exhibit the 2:1 female preponderance seen in MDD.[35-38] In contrast, patients with mid-life-onset dysthymia have a male preponderance, do not have comorbid major depressive episodes or family loading for depressive illness, and are relatively treatment-resistant.[27;35-38] Based on such data and our experience with older hypogonadal men, we and others have proposed[23;39-42] that a new-onset, low-grade chronic depressive condition develops in some middle-aged men; that this condition may be the psychiatric manifestation of

mid-life onset (and progressive) hypogonadism; and that testosterone replacement is an effective antidepressant in such men.

## 2.1 Objectives

Primary

To determine the efficacy of testosterone replacement therapy in inducing remission in middle-aged and older hypogonadal men with mid-life onset dysthymia.

Secondary

1. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in depressive symptom score ascertained using the **PHQ-9 scale** in men who meet the definition of mid-life dysthymia at baseline.

2. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in depressive symptom score using the **GDS-15 scale** in men who meet the definition of mid-life dysthymia at baseline.

3. To determine whether compared to placebo, testosterone replacement therapy is associated with a greater proportion of men who meet the definition of mid-life dysthymia at baseline who achieve remission at 12 months and maintain remission at 18 and 24 months

4. To determine whether compared to placebo, testosterone replacement therapy is associated with a greater proportion of men who achieve remission of mid-life-onset dysthymia at 6 months

5. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in depressive symptom score using the mood domain of the HIS-Q scale in men who meet the definition of mid-life dysthymia at baseline.

6. To determine the efficacy of testosterone replacement therapy, relative to placebo, in improving depressive symptoms assessed using mood domain of HIS-Q in middle-aged and older hypogonadal men, who have a depressive disorder at baseline defined by a PHQ-9 score of higher than 4.

7. To determine whether compared to placebo, testosterone replacement therapy is associated with greater improvements in depressive symptoms in all randomized subjects, assessed using the mood domain of HIS-Q.

8. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in sleep quality assessed using the sleep domain of the HIS-Q scale in all participants enrolled in the parent trial as well as in the subset enrolled in the PDD substudy

9. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in cognition assessed using the cognition domain of the HIS-Q scale in all participants enrolled in the parent trial as well as in the subset enrolled in the PDD substudy

10. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in the energy / vitality as measured by Energy Domain of the HIS-Q scale in all participants enrolled in the parent trial as well as in the subset enrolled in the PDD substudy

11. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in cognition assessed using the cognition domain of the HIS-Q scale in middle-aged and older hypogonadal men, who have a depressive disorder at baseline defined by a PHQ-9 score of higher than 4.

12. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in the energy / vitality as measured by Energy Domain of the HIS-Q scale in in middle-aged and older hypogonadal men, who have a depressive disorder at baseline defined by a PHQ-9 score of higher than 4.

13. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in Total and Free Testosterone, SHBG, DHT and estradiol in in middle-aged and older hypogonadal men enrolled in PDD sub-study

14. To determine whether compared to placebo, testosterone replacement therapy is associated with greater change from baseline in Total and Free Testosterone, SHBG,

DHT and estradiol in in middle-aged and older hypogonadal men, who have a depressive disorder at baseline defined by a PHQ-9 score of higher than 4.

15.

Exploratory:

To determine in all randomized subjects the frequency of patient reports of **adverse experiences** including:

- Hospitalizations for depressive or manic episodes
- Suicidality
- Other descriptors derived from MedDRA coding of psychiatric illnesses

Additionally, an exploratory analysis will characterize the group of randomized participants who meet the definition of a depressive disorder and who exhibit a response to TRT by improving their HIS-Q mood domain score more than the median.

## 2.2 Study Design

The TRAVERSE parent trial is a Phase 4, randomized, double-blind, placebo-controlled, multicenter study of topical TRT in symptomatic hypogonadal men with increased risk for CV disease. The initial planned study enrollment is approximately 6,000 subjects based on the projected timing when 256 MACE will occur under initial assumptions of the annual event rate, subject accrual rate, and study discontinuation rate. There will be approximately 400 sites in North America and possibly Puerto Rico. An Interactive Response Technology (IRT) system will randomize subjects to receive either topical testosterone or placebo in a 1:1 ratio. Randomization will be stratified by pre-existing CV disease (Yes/No). Titration of testosterone dose will occur in subjects receiving active testosterone, while sham dosage titrations will occur in subjects receiving placebo gel via the central IRT system. The Screening Period is up to 50 days prior to first study drug dose. Once subjects meet all of the eligibility criteria during Screening, they will be randomized (1:1 ratio) to active study drug or placebo and will be followed until the study ends. Importantly, randomized subjects who elect to discontinue study

drug will also be followed until the study ends unless the subject dies or withdraws from the study completely (withdrawal of informed consent) earlier. Subjects who discontinue study drug will still be asked to follow their regularly scheduled protocol visits. Subjects who interrupt study drug will be allowed to restart study drug at any time.

The aim of the TRAVERSE-PDD sub-study is to determine the benefits on depressive symptoms of TRT in middle-aged men. The study will compare those randomized to testosterone and those randomized to control among those eligible for the TRAVERSE-PDD sub-study. The sub-study will be nested within the parent trial. Among the 6,000 men in the parent study, those who meet the additional eligibility criteria and sign informed consent for this "TRAVERSE-PDD sub-study" will be enrolled in this sub-study.

Participants eligible for PDD sub-study must fulfill all of the following inclusion criteria (with the exception of the analysis using HIS-Q mood score which will be performed on all participants in the parent trial):

- Willingness to participate in the PDD sub-study

and PHQ-9 criteria from SV2 visit:

- Total PHQ-9 score > 4
- PHQ-9 Item 2 score of 1 or 2
- PHQ-9 Item 10 score of > 0

Additionally, men must not meet any of the below PHQ-9 criteria from visit SV2:

- Total PHQ-9 score > 15
- PHQ-9 Items 1 or 2 score of 3
- PHQ-9 Item 9 score of > 0

ePRO tablet programming will identify men that meet the above preliminary PHQ-9 depression sub-study eligibility criteria from SV2. At the Baseline Visit, these men will then be asked additional eligibility questions (Appendix N from study Protocol) to assess their eligibility for the PDD sub-study via the PRO tablet based on questions (Appendix N from study Protocol) adapted from the diagnosis of PDD as described in Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria for PDD with low-grade symptoms.

Finally, men who are still eligible via the above criteria will be asked to take the GDS-15 questionnaire. Men with GDS-15 scores between 5 – 9 will then be entered into the PDD sub-

study.
If subjects respond to any of the above questions in a way that makes them ineligible for the PDD sub-study, the tablet will skip the remainder of the questions related to PDD sub-study eligibility.

*Participants must not meet any of the following exclusion criteria to be enrolled in this sub-study:*

- Lifetime history of any major depressive episode
- Onset of any mood disorder prior to age 40 years
- History of treatment for a mood disorder longer than 6 months (e.g., antidepressant medication, psychotherapy)
- History of a suicide attempt

### 2.2.1 Variables Used for Stratification of Randomization

Randomization will descend from the parent trial; no additional randomization or stratification will be imposed in this sub-study. In the parent trial, randomization will be stratified by pre-existing CV disease (Yes/No). It is expected that in the parent trial 30% of the randomized subjects will satisfy inclusion criteria for pre-existing CV disease criteria (secondary prevention), and the remaining 70% will satisfy CV risk factors criteria (primary prevention) combined. Analyses in this PDD sub-study will acknowledge stratified randomization.

## 2.3 Endpoints

### 2.3.1 Primary Efficacy Endpoint

The primary efficacy outcome of the TRAVERSE-PDD is the proportion of men who achieve **remission of mid-life-onset dysthymia at the 12-month visit.**

The remission of low-grade, mid-life-onset dysthymia will be defined by the following criteria:

a) GDS-15 <5, AND

b) An answer of “no” to the query "Give your best guess: Over the past 6 months, have you been feeling sad or depressed more days than not, even if you felt okay sometimes?"

### 2.3.2 Secondary Efficacy Endpoint

Secondary endpoints include:

- Change from baseline in depressive symptom score ascertained using the **PHQ-9 scale** in men who meet the definition of mid-life dysthymia at baseline;
- Change from baseline in depressive symptom score using the **GDS-15 scale** in men who meet the definition of mid-life dysthymia at baseline;
- Change from baseline in depressive symptom score using the **mood domain of the HIS-Q scale** in men who meet the definition of mid-life dysthymia at baseline;
- Proportion of men who meet the definition of mid-life dysthymia at baseline who achieve remission at 12 months and maintain remission at 18 and 24 months
- Proportion of men who achieve remission of mid-life-onset dysthymia at 6 months
- Change from baseline in depressive symptoms ascertained using the HIS-Q score (mood domain) in all participants enrolled in the parent trial who meet the definition of a depressive disorder (screening or baseline PHQ-9 score 5 or greater)
- Change from baseline in depressive symptoms ascertained using the HIS-Q score (mood domain) in all participants enrolled in the parent trial
- Change from baseline in sleep assessed using the sleep domain of the HIS-Q scale in all participants enrolled in the parent trial as well as in the subset enrolled in the PDD substudy
- Change from baseline in cognition assessed using the cognition domain of the HIS-Q scale in all participants enrolled in the parent trial as well as in the subset enrolled in the PDD substudy
- Change from baseline in the energy / vitality as measured by Energy Domain of the HIS-Q scale in all participants enrolled in the parent trial as well as in the subset enrolled in the PDD substudy
- Change from baseline in cognition assessed using the cognition domain of the HIS-Q scale in all enrolled men, who have a depressive disorder at baseline defined by a PHQ-9 score of higher than 4.

- Change from baseline in the energy / vitality as measured by Energy Domain of the HIS-Q scale in all enrolled men, who have a depressive disorder at baseline defined by a PHQ-9 score of higher than 4.

### 2.3.3 Exploratory Endpoints

Exploratory endpoints will include:

- patient reports of **adverse experiences relevant to sub-study endpoints**, including
    - Hospitalizations for depressive or manic episodes
    - Suicidality
    - Other descriptors derived from MedDRA coding of psychiatric illnesses

### 2.3.4 Safety Endpoint

Safety assessments will be incorporated into the parent trial. No additional safety endpoints will be specified in this sub-study.

## 2.4 Sample Size Justification

The intended sample will include all eligible individuals meeting inclusion / exclusion criteria for the sub-study. We provide computations below demonstrating that **the anticipated sample size will provide 90% power to detect a remission risk difference of at least 0.11 between the two randomized arms under reasonable assumptions, assuming type-I error rate 0.05**. This is a subtler difference than anticipated per existing data and clinical experience. The planned sample size will therefore be adequate to the needs of the sub-study.

Population-based studies suggest that 1.5-10% of the general middle-aged and elderly male population is affected by PDD.[27;28;43-46] The prevalence of dysthymia in middle-aged and older hypogonadal men is likely substantially higher. In TRiUS, a multicenter, 12-month observational registry (N=849) of hypogonadal men, 52% of men enrolled had mild to moderate depression, defined by Patient Health Questionnaire-9 scores of 5-14.[47]

Based on these data, we anticipate that among men age 45-80 with rigorously documented hypogonadism, the prevalence of mid-life-onset, low-grade persistent depression will be at least

40%.[47;48] However, exclusion criteria are likely to remove some of these individuals from the population eligible from the parent TRAVERSE trial. Assuming conservatively that half of the men are excluded by these criteria or decline to participate in the sub-study, we anticipate at least 20% of 6000 men, or 1200 men eligible and willing to participate in the TRAVERSE-PDD sub-study.

There are two relatively small randomized, placebo-controlled clinical trials of testosterone replacement in the treatment of low-grade depression in middle-aged and older men. Shores et al[24] enrolled 33 hypogonadal men (total testosterone level ≤ 280 ng/dL), aged 50 years or older with "subthreshold depression" (dysthymia or minor depression, according to DSM-IV) to a 12-week trial of testosterone replacement; the remission rate of those randomized to testosterone vs placebo was 53% vs. 19% (p = .04), a nominal risk ratio of approximately 2.8. Seidman et al[23] randomized a similar group (N=23), with mid-life-onset dysthymia and testosterone < 350ng/dL) to testosterone vs placebo; remission rate of those randomized to testosterone vs placebo was 54% vs. 10% (p = .03), a nominal risk ratio greater than 5.

Based these data and prior clinical experience, we more conservatively hypothesize that 40% of such men will respond to testosterone replacement with remission of depressive symptoms, as compared to 20% receiving placebo, an absolute difference in risk of 20% and a nominal risk ratio of 2.

The primary comparison of interest will be the proportion of individuals in the TRAVERSE-PDD substudy who express remission of PDD at one year in the testosterone arm vs. the placebo arm. We anticipate cumulative remission rate of 20% among men randomized to the placebo arm. Finally, we assume 15% cumulative missingness on depression endpoints at 1 year due to death, loss to follow-up for all other reasons, and other sources of missing data at one year.

Under these assumptions, and assuming type-I error rate 0.05, we will have more than 95% power to detect a difference in remission rates of 20% between TRT and placebo at 1 year if 780 participants were enrolled in the sub-study. The same sample-size would yield approximately 90% power to detect a difference in rates as low as 11% and 85% power to detect a difference in rates as low as 10%, which we consider to approximate the lower bound on a clinically significant effect. We therefore anticipate adequate power to detect clinically meaningful effects even with substantially lesser enrollment than anticipated.

### 2.5 Interim Analysis

No interim analyses are planned.

### 2.6 Multiplicity Testing Procedures for Type-I Error Control

As hypotheses are prespecified and mutually reinforcing, no multiplicity adjustments are planned.

## 3.0 Analysis Populations and Important Subgroups

### 3.1 Analysis Populations

The **primary efficacy population** will be a subset of the Efficacy Analysis Set (EAS) for the parent TRAVERSE Trial (subjects analyzed according to their treatment assignment at randomization), and it will include participants who were enrolled to this sub-study. Participants will be grouped by their randomized treatment assignment. Eligible subjects who satisfy the inclusion and exclusion criteria for this sub-study will be included in analyses. Data from eligible measurements will be included, where eligible is defined as being obtained from questionnaires or scores for which at least 80% of items are non-missing.

**Efficacy Analysis Set (EAS)** for the parent TRAVERSE Trial will be used in the analysis of the secondary endpoints (HIS-Q mood domain and other endpoints).

For any given analysis, subjects with an eligible baseline measurement and at least one requisite eligible post-baseline measurement will be included.
A **modified efficacy population** maybe examined in limited secondary analyses. For these analyses, participants will be censored at any time (and subsequently) that they are deemed treatment noncompliant if they are withdrawn from treatment.

### 3.2 Pre-specified subgroups

Trial subpopulations given specific consideration will include categorizations by:

- Self-identified race and ethnicity if subgroup sample sizes will support this
- Age (45-65 vs. 65+)

- Prior CV status (a design feature)
- Baseline testosterone level (< 250 vs. 250+ ng/dl)

### 3.3 Definition of Treatment Groups

Treatment groups will be defined according to randomization and stratification in the parent trial as defined above.

## 4.0 Analysis Conventions

**The Operating Procedures of Data Collection and Analysis.** In case the following data operating procedures do not cover any particular aspect of data analysis or variable definitions in this sub-study Statistical Analysis Plan, the operating procedures of the Parent TRAVERSE Trial will be binding. For participants with duplicate ID in the database, data from the duplicated subjects will be excluded from the corresponding analysis sets in this sub-study. An average number of days in a year will be set as 365.25, and an average number of days in 1 month will be set as 365.25/12 for exposure calculation and time to event analyses. The pre-existing cardiovascular disease will be defined, in accordance with parent trial definition, as an occurrence of one or more of the three conditions: coronary artery disease, cerebrovascular disease or peripheral arterial disease. If the same questionnaire data is collected more than once on the same day, the worst score (the highest or the lowest one, depending on the questionnaire) will be used from that day. If there are multiple assessments around the nominal day, the assessment closest to the nominal day will be used.

### 4.1 Definition of Baseline

Baseline on each outcome measure will be defined as the last available measurement obtained prior to the first dose of study drug (defined as on or before Day 1) (Protocol Appendix C).

### 4.2 Definition of Final Observation

Final observation for each analysis will be the final assessment affiliated with the relevant visit. For example, for twelve-month assessment of PDD, the final observation of each relevant measurement affiliated with the 12-month visit will be included in analyses.

### 4.3 Definition of Visit Windows

Definitions of the visit windows (baseline and on-treatment) are presented in Section 11.0 (Tables 11.1-12.3) and schedule of sub-study activities is presented in Section 10.0.

## 5.0 Demographics, Baseline Characteristics, Medical History, and Previous and Concomitant Medications

### 5.1 Baseline Characteristics and participant disposition

Demographic assessments will be obtained from the parent TRAVERSE trial and will include at minimum age, race, ethnicity, height, weight, body mass index, vital signs and laboratory assessments at baseline. Other characterizations may be considered as appropriate.

Statistics for continuous variables will include mean, median, standard deviation, minimum, maximum, and sample size for each treatment group. Binary variables will be described with frequencies and percentages.

### 5.2 Medical History

Medical history will likewise be obtained in the parent TRAVERSE trial, and will include at minimum history of depression, PDD and mood disorders; CV history; nicotine and alcohol use.

### 5.3 Previous Treatment and Concomitant Medications

Prior and concomitant medication use will be derived from assessments in the parent trial and considered in this sub-study as appropriate.

### 5.4 Study Drug Exposure and Compliance

In the parent trial, the bottle weight of the returned study drug bottles will be compared to an average bottle weight to determine compliance. The number of doses that should have occurred

between visits will be calculated and compliance computed as a fraction of this dose, and reported for each group in this sub-study at one and two years post-baseline.

Total patient-years of exposure will be calculated by summing the duration of treatment, defined as last dose date – first dose date + 1 (in days), for all subjects in the analysis set and dividing this sum by 365.25 (= 1 year). In addition, the number and percentage of subjects exposed to study drug will be summarized for the following categories of exposure duration: ≤ 1 month, > 1 to 2 months, > 2 to 3 months, > 3 to 6 months, > 6 months to 1 year, > 1 to 2 years, etc.

# 6.0 Analysis of Endpoints

## 6.1 General Considerations

Analyses will be supported by descriptive comparisons of the two treatment arms. Primary and secondary efficacy endpoints will be analyzed using simple and multiple regression modeling approaches. In all cases, treatment effects will be accompanied by robust variance estimates and confidence intervals derived thereby. All analyses will be adjusted for the effect of prior CV event (a stratifying factor in the TRAVERSE trial randomization).

### 6.1.1 Primary Endpoint

The multiplicative difference in the risk (i.e. risk ratio) of **PDD remission in TRT relative to control at 12 months** will be estimated in a mixed model regression via the modified Poisson method.[50] Secondary analyses controlling for age and other covariates (Section 3.2) will also be presented. A sensitivity analysis may consider those individuals who express remission at both follow-up time points 6 and 12 months.

## 6.2 Secondary Endpoints

A between-group comparison will be performed to assess rates of **participants who maintained PDD remission status at the 24-month evaluation** using the same approach as described for the primary comparison.

To model **change in PDD symptom scores and hormone levels with time**, we will use mixed-effects linear regression analysis with random intercept. The model will incorporate effects for visit and visit-by-treatment interaction. The average difference between baseline and treatment

over the entire treatment period will be obtained using a treatment contrast and associated confidence interval. The statistical significance of the treatment difference will be obtained using this interval and an omnibus F-test of the visit-by-treatment interaction term over all follow-up times. An unstructured covariance matrix will be assumed; where appropriate or necessary (e.g. due to lack of convergence of the unstructured covariance model), the compound symmetry assumption will be considered.

To assess difference between proportions of men who experience **remission of low grade PDD in placebo vs. testosterone arm at specific time points**, non-linear mixed model regression model (repeated measures modified Poisson regression) will be used.[51] The model will incorporate effects for visit, treatment and visit-by-treatment interaction. The average difference between baseline and treatment over the entire treatment period will be obtained using a treatment contrast and associated confidence interval. The statistical significance of the treatment difference will be obtained using this interval and an omnibus F-test of the visit-by-treatment interaction term over all follow-up times. An unstructured covariance matrix will be assumed; where appropriate or necessary (e.g. due to lack of convergence of the unstructured covariance model), the compound symmetry assumption will be considered.

**Proportion of individuals experiencing remission at 6 months** will be considered in an independent analysis following the plan described above.

Change from baseline in HIS-Q score (including mood and other domains) over time will be analyzed in study participants who meet the criteria for a depressive disorder (screening or baseline PHQ-9 score > 4) using mixed-effects linear regression analysis as described above.

Additionally, change from baseline in HIS-Q score (including mood and other domains) over time will be analyzed in similar fashion as other continuous endpoints in all randomized participants from the parent trial. Analyses for other secondary endpoints will be conducted in parallel fashion.

### 6.3 Exploratory analyses

Association studies tracking **contemporaneous change in outcomes** will be conducted using linear and generalized linear analyses, and will include:

- Relationship between changes in mood and sexual function scores derived from the sexual function sub-study
- Relationship between PDD remission and reports of energy and vitality embedded in the HIS-Q (Energy and Mood domains)
- Change in T levels in blood relative to baseline and change in endpoints

## 7.0 Missing Data

There are no plans to impute missing data in this substudy. In the event that such is required by a regulatory body or journal editor, multiple imputation of endpoints and covariates by the MICE methodology[52, 53] will be considered as appropriate. This method is notable for being able to handle clustering of repeated measures at the participant level, a feature of the design of this trial and sub-study.

## 8.0 Changes in this version

Clarified definition of primary and secondary analysis samples. Clarified operating procedures of data collection and definition.

## 9.0 References

1. Ebinger M, Sievers C, Ivan D, et al. Is there a neuroendocrinological rationale for testosterone as a therapeutic option in depression? J Psychopharmacol 2008;

2. Seidman SN, Araujo AB, Roose SP, et al. Low testosterone levels in elderly men with dysthymic disorder. Am J Psychiatry 2002;159: 456-459

3. Markianos M, Tripodianakis J, Sarantidis D, et al. Plasma testosterone and dehydroepiandrosterone sulfate in male and female patients with dysthymic disorder. J Affect Disord 2007;101: 255-258

4. Shores MM, Sloan KL, Matsumoto AM, et al. Increased incidence of diagnosed depressive illness in hypogonadal older men. Arch Gen Psychiatry 2004;61: 162-167

5. Almeida OP, Yeap BB, Hankey GJ, et al. Low free testosterone concentration as a potentially treatable cause of depressive symptoms in older men. Arch Gen Psychiatry 2008;65: 283-289

6. Perry PJ, Yates WR, Williams RD, et al. Testosterone therapy in late-life major depression in males. J Clin Psychiatry 2002;63: 1096-1101

7. Nian Y, Ding M, Hu S, et al. Testosterone replacement therapy improves health-related quality of life for patients with late-onset hypogonadism: a meta-analysis of randomized controlled trials. Andrologia 2016;

8. Ahern T, Swiecicka A, Eendebak RJ, et al. Natural history, risk factors and clinical features of primary hypogonadism in ageing men: Longitudinal Data from the European Male Ageing Study. Clin Endocrinol (Oxf) 2016;

9. Seidman SN. Testosterone deficiency and mood in aging men: pathogenic and therapeutic interactions. World J Biol Psychiatry 2003;4: 14-20

10. Wang C, Swerdloff RS, Iranmanesh A, et al. Transdermal testosterone gel improves sexual function, mood, muscle strength, and body composition parameters in hypogonadal men. Testosterone Gel Study Group. J Clin Endocrinol Metab 2000;85: 2839-2853

11. Anderson RA, Bancroft J, Wu FC. The effects of exogenous testosterone on sexuality and mood of normal men. J Clin Endocrinol Metab 1992;75: 1503-1507

12. Burris AS, Banks SM, Carter CS, et al. A long-term, prospective study of the physiologic and behavioral effects of hormone replacement in untreated hypogonadal men. J Androl 1992;13: 297-304

13. Luisi M, Franchi F. Double-blind group comparative study of testosterone undecanoate and mesterolone in hypogonadal male patients. J Endocrinol Invest 1980;3: 305-308

14. McNicholas TA, Dean JD, Mulder H, et al. A novel testosterone gel formulation normalizes androgen levels in hypogonadal men, with improvements in body composition and sexual function. BJU Int 2003;91: 69-74

15. Steidle C, Schwartz S, Jacoby K, et al. AA2500 testosterone gel normalizes androgen levels in aging males with improvements in body composition and sexual function. J Clin Endocrinol Metab 2003;88: 2673-2681

16. Kunelius P, Lukkarinen O, Hannuksela ML, et al. The effects of transdermal dihydrotestosterone in the aging male: a prospective, randomized, double blind study. J Clin Endocrinol Metab 2002;87: 1467-1472

17. Wang C, Cunningham G, Dobs A, et al. Long-term testosterone gel (AndroGel) treatment maintains beneficial effects on sexual function and mood, lean and fat mass, and bone mineral density in hypogonadal men. J Clin Endocrinol Metab 2004;89: 2085-2098

18. Schiavi RC, White D, Mandeli J, et al. Effect of testosterone administration on sexual behavior and mood in men with erectile dysfunction. Arch Sex Behav 1997;26: 231-241

19. Seidman SN, Spatz E, Rizzo C, et al. Testosterone replacement therapy for hypogonadal men with major depressive disorder: a randomized, placebo-controlled clinical trial. J Clin Psychiatry 2001;62: 406-412

20. Orengo CA, Fullerton L, Kunik ME. Safety and efficacy of testosterone gel 1% augmentation in depressed men with partial response to antidepressant therapy. J Geriatr Psychiatry Neurol 2005;18: 20-24

21. Seidman SN, Miyazaki M, Roose SP. Intramuscular testosterone supplementation to selective serotonin reuptake inhibitor in treatment-resistant depressed men: randomized placebo-controlled clinical trial. J Clin Psychopharmacol 2005;25: 584-588

22. Pope HG, Jr., Amiaz R, Brennan BP, et al. Parallel-group placebo-controlled trial of testosterone gel in men with major depressive disorder displaying an incomplete response to standard antidepressant treatment. J Clin Psychopharmacol 2010;30: 126-134

23. Seidman SN, Orr G, Raviv G, et al. Effects of testosterone replacement in middle-aged men with dysthymia: a randomized, placebo-controlled clinical trial. J Clin Psychopharmacol 2009;29: 216-221

24. Shores MM, Kivlahan DR, Sadak TI, et al. A randomized, double-blind, placebo-controlled study of testosterone treatment in hypogonadal older men with subthreshold depression (dysthymia or minor depression). J Clin Psychiatry 2009;70: 1009-1016

25. Bloch M, Schmidt PJ, Danaceau MA, et al. Dehydroepiandrosterone treatment of midlife dysthymia [see comments]. Biol Psychiatry 1999;45: 1533-1541

26. Svanborg C, Wistedt AA, Svanborg P. Long-term outcome of patients with dysthymia and panic disorder: a naturalistic 9-year follow-up study. Nord J Psychiatry 2008;62: 17-24

27. Devanand DP. Dysthymic disorder in the elderly population. Int Psychogeriatr 2014;26: 39-48

28. Murphy JA, Byrne GJ. Prevalence and correlates of the proposed DSM-5 diagnosis of Chronic Depressive Disorder. J Affect Disord 2012;139: 172-180

29. Rhebergen D, Graham R. The re-labelling of dysthymic disorder to persistent depressive disorder in DSM-5: old wine in new bottles? Curr Opin Psychiatry 2014;27: 27-31

30. Baune BT, Adrian I, Arolt V, et al. Associations between major depression, bipolar disorders, dysthymia and cardiovascular diseases in the general adult population. Psychother Psychosom 2006;75: 319-326

31. Singh-Manoux A, Akbaraly TN, Marmot M, et al. Persistent depressive symptoms and cognitive function in late midlife: the Whitehall II study. J Clin Psychiatry 2010;71: 1379-1385

32. Tully PJ, Baumeister H, Martin S, et al. Elucidating the Biological Mechanisms Linking Depressive Symptoms With Type 2 Diabetes in Men: The Longitudinal Effects of Inflammation, Microvascular Dysfunction, and Testosterone. Psychosom Med 2016;78: 221-232

33. Seidman SN. Neuroendocrinology of Mood Disorders. In: Stein DJ, Schatzberg AF, Kupfer DJ, eds. Textbook of Mood Disorders 1st ed. Washington, DC: American Psychiatric Press; 2005:

34. Huhtaniemi I. Late-onset hypogonadism: current concepts and controversies of pathogenesis, diagnosis and treatment. Asian J Androl 2014;16: 192-202

35. Devanand DP, Nobler MS, Singer T, et al. Is dysthymia a different disorder in the elderly? Am J Psychiatry 1994;151: 1592-1599

36. Kocsis JH. Geriatric dysthymia. J Clin Psychiatry 1998;59 Suppl 10: 13-15

37. Pinquart M, Duberstein PR, Lyness JM. Treatments for later-life depressive conditions: a meta-analytic comparison of pharmacotherapy and psychotherapy. Am J Psychiatry 2006;163: 1493-1501

38. Bartels SJ, Dums AR, Oxman TE, et al. Evidence-Based Practices in Geriatric Mental Health Care. Psychiatr Serv 2002;53: 1419-1431

39. Kanayama G, Amiaz R, Seidman S, et al. Testosterone supplementation for depressed men: current research and suggested treatment guidelines. Exp Clin Psychopharmacol 2007;15: 529-538

40. Amiaz R, Seidman SN. Testosterone and depression in men. Curr Opin Endocrinol Diabetes Obes 2008;15: 278-283

41. Seidman SN. Androgens and the aging male. Psychopharmacol Bull 2007;40: 205-218

42. Seidman SN, Weiser M. Testosterone and mood in aging men. Psychiatr Clin North Am 2013;36: 177-182

43. Beekman AT, Deeg DJ, Smit JH, et al. Dysthymia in later life: a study in the community. J Affect Disord 2004;81: 191-199

44. Steffens DC, Skoog I, Norton MC, et al. Prevalence of depression and its treatment in an elderly population: the Cache County study. Arch Gen Psychiatry 2000;57: 601-607

45. Costa E, Barreto SM, Uchoa E, et al. Prevalence of International Classification of Diseases, 10th Revision common mental disorders in the elderly in a Brazilian community: The Bambui Health Ageing Study. Am J Geriatr Psychiatry 2007;15: 17-27

46. Karg RS, Bose J, Batts KR, et al. Past Year Mental Disorders among Adults in the United States: Results from the 2008-2012 Mental Health Surveillance Study. 2012;

47. Khera M, Bhattacharya RK, Blick G, et al. The effect of testosterone supplementation on depression symptoms in hypogonadal men from the Testim Registry in the US (TRiUS). Aging Male 2012;15: 14-21

48. Huhtaniemi IT. Andropause--lessons from the European Male Ageing Study. Ann Endocrinol (Paris) 2014;75: 128-131

49. Snyder PJ, Bhasin S, Cunningham GR, Matsumoto AM, Stephens-Shields AJ, Cauley JA, Gill TM, Barrett-Connor E, Swerdloff RS, Wang C, Ensrud KE, Lewis CE, Farrar JT, Cella D, Rosen RC, Pahor M, Crandall JP, Molitch ME, Cifelli D, Dougar D, Fluharty L, Resnick SM, Storer TW, Anton S, Basaria S, Diem SJ, Hou X, Mohler ER 3rd, Parsons JK, Wenger NK, Zeldow B, Landis JR, Ellenberg SS; Testosterone Trials Investigators. Effects of Testosterone Treatment in Older Men. N Engl J Med. 2016 Feb 18;374(7):611-24. doi: 10.1056/NEJMoa1506119. PubMed PMID: 26886521; PubMed Central PMCID: PMC5209754.

50. Guangyong Zou; A Modified Poisson Regression Approach to Prospective Studies with Binary Data, American Journal of Epidemiology, Volume 159, Issue 7, 1 April 2004, Pages 702–706, https://doi.org/10.1093/aje/kwh090

51. Ma Y, Mazumdar M, Memtsoudis SG. Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. Reg Anesth Pain Med 2012; 37: 99–105.

52. Van Buuren, S., Flexible Imputation of Missing Data. 2012, Chapman & Hall/CRC.

53. Van Buuren, S. and K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 2011. 45(3): p. 67.

# 10.0 Schedule of Activities

Study activities from main protocol

| Activity | SV2 | D1 (Baseline) | M 6 W 26 | M 12 W 52 (year 1) | M 18 W 78 | M 24 W 104 (year 2) | M 60 W 260/FV(Year 5) | PD |
|---|---|---|---|---|---|---|---|---|
| PHQ-9[a] | X | | | | | | | |
| PGI-I Hypogonadism | | | X | | | X | X | X |

[a]PHQ-9 score ranges from 0 to 27 (lower values better).

Additional study activities specific for PDD sub-study

| Activity | D1 | M 6 W26 | M 12 W 52 | M 18 W 78 | M 24 W 104 | M 60 W 260/FV (Year 5) | PD |
|---|---|---|---|---|---|---|---|
| PHQ-9[a] | | X | X | X | X | | |
| Additional Persistent Depressive Disorder (PDD) Questions | X | X | X | | X | | |
| GDS[b] | X | X | X | | X | | |
| PGI-I Mood | | | X | | X | X | X |

[a]PHQ-9 score ranges from 0 to 27 (lower values better).

[b]GDS score ranges from 0 to 15 (lower values better).

# 11.0 Efficacy Analysis Time Windows

## 11.1 For Activity: PHQ-9

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M6 | 182 | 92 - 273 |
| M12 | 364 | 274 - 455 |
| M18 | 546 | 456 - 637 |
| M24 | 728 | 638 - 1093 |

| | | |
|---|---|---|
| Final Visit | | 2 to ≤ 2 days after the last dose of study drug |

## 11.2 For Activity: Additional PDD Questions, GDS

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M6 | 182 | 92 - 273 |
| M12 | 364 | 274 - 546 |
| M24 | 728 | 547 - 1093 |
| Final Visit | | 2 to ≤ 2 days after the last dose of study drug |

## 11.3 For Activity: PGI-I Mood

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤1 |
| M12 | 364 | 274 - 546 |
| M24 | 728 | 547 - 1093 |
| M60 | 1820 | 1460-2180 |
| Final Visit / PD | | 2 to ≤ 2 days after the last dose of study drug |

# Statistical Analysis Plan

# Supplement for Sexual Function Efficacy Sub-Study

## Study M16-100

## Testosterone Replacement Therapy for Assessment of Long-Term Vascular Events and Efficacy ResponSE in Hypogonadal Men (TRAVERSE) Study

## Date: 01 January 2023

## Version 3.4

## Table of Contents

# List of Abbreviations

| | |
|---|---|
| ANCOVA | Analysis of covariance |
| CV | Cardiovascular |
| DISF-SRII | DeRogatis Interview for Sexual Function – Self Report |
| ESC | Executive Steering Committee |
| FAS | Full Analysis Set |
| HIS-Q | Hypogonadism Impact of Symptoms Questionnaire |
| IIEF | International Index of Erectile Function |
| I-PSS | International Prostate Symptom Score |
| IRT | Interactive Response Technology |
| MACE | Major Adverse Cardiac Event |
| MICE | Multiple Imputation by Chained Equations |
| MID | Minimal Important Difference |
| PDQ | Psychosexual Daily Questionnaire |
| PGI-I | Patient Global Index of Improvement |
| SAP | Statistical analysis plan |
| TRT | Testosterone Replacement Therapy |

# 1.0 Introduction

This statistical analysis plan supplement (SAP) provides details to elaborate statistical methods for data collected as outlined in the protocol (Amendment 1 dated 26 February 2018[1]) for efficacy sub-study of Study M16-100 parent trial and describes analysis conventions to guide the statistical programming work. The scope of this SAP is limited to only the sexual function efficacy sub-study.

All analyses will be performed using SAS Version 9.2 or higher (SAS Institute, Inc., Cary, NC 27513) under the UNIX operating system. The SAP will be signed off before the study database is locked.

This SAP will *not* be updated in case of future (administrative or minor) amendments to the protocol unless the changes have any impact on the analysis of study data described here.

# 2.0 Study Background

## 2.1 Objective

**Primary Objective:**

- Determine whether TRT for 1 year improves overall sexual activity, using Question 4 of the PDQ, more than placebo.

**Secondary Objectives:**

- Determine whether TRT for 1 year improves sexual desire, using the libido domain of HIS-Q instrument, more than placebo.
- Determine whether, compred to placebo, TRT for 1 year improves hypogonadal symptoms, assessed using the HIS-Q instrument
- Determine whether, compred to placebo, TRT for 1 year improves sexual symptoms of hypogonadism asssessed using the sexual domain of the HIS-Q instrument

- Determine whether TRT for 1 year improves erectile function, using the erectile function domain of the IIEF, more than placebo.
- Determine whether improvements in sexual activity, sexual desire, and erectile function over 1 year are maintained after 2 years of TRT versus placebo.
- Determine whether TRT for 1 year improves PDQ Question 4 score by clinically meaningful change (≥ 0.6), more than placebo.
- Determine whether TRT improves sexual desire over the course of the 3-year follow-up, using the libido domain of HIS-Q instrument, more than placebo.
- Determine whether, compared to placebo, TRT improves hypogonadal symptoms assessed using the HIS-Q instrument over the course of the 3-year follow-up
- Determine whether, compared to placebo, TRT improves sexual symptoms of hypogonadism assessed using the HIS-Q instrument over the course of the 3-year follow-up
- Determine the distribution of hypogonadism symptoms at baseline ((a) decreased sexual desire or libido, (b) decreased spontaneous erections, (c) decreased energy or fatigue/feeling tired, (d) low mood or depressed mood, (e) loss of body axillary and pubic hair or (f) reduced shaving or hot flashes) among all men enrolled in the parent trial

**Exploratory Objectives:**

- Compare the effect of TRT versus placebo therapy on patient global impression of change in sexual function using a patient global impression of change (PGI-I Libido) question.
- Evaluation of minimally important difference (MID) for the score obtained from HIS-Q Libido question.
- Evaluation of the relationship between changes in testosterone levels and changes in sexual function outcomes at year 1, in patients randomized to testosterone arm.

## 2.2 Study Design

### 2.2.1 Study Design and Design Diagram

The TRAVERSE parent trial is a Phase 4, randomized, double-blind, placebo-controlled, multicenter study of topical TRT in symptomatic hypogonadal men with increased risk for CV disease. The initial planned study enrollment is approximately 6,000 subjects based on the projected timing when 256 MACE will occur under initial assumptions of the annual event rate, subject accrual rate, and study discontinuation rate. There will be approximately 400 sites in North America and possibly Puerto Rico. An Interactive Response Technology (IRT) system will randomize subjects to receive either topical testosterone or placebo in a 1:1 ratio. Randomization will be stratified by pre-existing CV disease (Yes/No). Titration of testosterone dose will occur in subjects receiving active testosterone, while sham dosage titrations will occur in subjects receiving placebo gel via the non-blinded central IRT system. The Screening Period is up to 50 days prior to first dose of study drug. Once subjects meet all of the eligibility criteria during Screening, they will be randomized (1:1 ratio) to active study drug or placebo and will be followed until the study ends. Importantly, randomized subjects who elect to discontinue study drug will also be followed until the study ends unless the subject withdraws from the study completely (withdrawal of informed consent). Subjects who discontinue study drug will still be asked to follow their regularly scheduled protocol visits. Subjects who interrupt study drug will be allowed to restart study drug at any time.

The aim of this sexual function sub-study is to determine the benefits on sexual function of TRT in middle-aged men, and whether the benefits in older men observed previously are durable over time and whether these benefits can be seen in men with CV disease and CV risk who are at higher risk of having sexual symptoms.

Accordingly, the sexual function sub-study will determine the efficacy of TRT in improving sexual activity and function in middle-aged and older hypogonadal men with low libido, as indicated by DISF-SRII Section I score $\leq 20$ at baseline, who have increased CV risk and will determine the durability of treatment effect beyond the first year of intervention.

### 2.2.2 Variables Used for Stratification at Randomization

Randomization will be stratified by pre-existing CV disease (Yes/No). It is expected that in the parent trial 30% of the randomized subjects will satisfy inclusion criteria for pre-existing CV disease criteria (secondary prevention), and the remaining 70% will satisfy other CV risk factors criteria (primary prevention) combined; these proportions will be monitored by the study team.

The ESC and Sponsor may decide to cap the secondary prevention cohort if that cohort is found to consistently exceed 70% of the total population enrolled or if the pooled primary event rate falls below projections.

## 2.3 Sexual Function Efficacy Sub-Study Endpoints

### 2.3.1 Questionnaires and Rating Scales for Sexual Function Endpoints

Schedule of measurements is provided in Section 9.

- **International Prostate Symptom Score-1 (I-PSS)**
  The I-PSS is a questionnaire (example Appendix F in the Protocol) used to help assess urination patterns and define the severity of BPH or lower urinary tract symptoms.[2] The standardized I-PSS instrument will be administered to the subject at designated visits. Subjects will complete the I-PSS as indicated in Appendix C in the study Protocol[1].

- **Sexual Function Questionnaires DeRogatis Interview for Sexual Function – Male (DISF – SRII)©**
  The DISF-SRII[3] is a brief inventory of questions about sexual thoughts and activity. The questions are divided into 5 sections that ask about different aspects of the individual's sexual experiences. Some questions request answers in terms of "how often" one engages in certain sexual activities. Other questions ask "how intense" one's sexual experiences are. A third type of question asks how much one "enjoyed" or was "satisfied" by different aspects of sexual activities and relationship. Each section has one or more scale definition boxes located alongside the questions. For Section I of the instrument, the scales range from 0 – 7 or 0 – 5. Results from Section

I (Sexual Desire domain) will determine whether or not a subject is included in the sexual function sub-study analysis set. Section I of the DISF - SRII will only be asked at Baseline. Subjects will complete the DISF – SRII Section I as indicated in Appendix C (Main Study) and Appendix O (Sub-study). The example DISF-SRII is provided in Appendix G in the Protocol[1].

- **Hypogonadism Impact of Symptoms Questionnaire (HIS – Q)**
  The HIS-Q is a validated patient-reported outcome measurement to evaluate the symptoms of hypogonadism and assess changes in symptoms among men with hypogonadism in response to treatment with TRT.[4] It includes 28 item questions assessing 5 domains (sexual, energy, sleep, cognition and mood) and 2 sexual subdomains (libido and sexual function): sexual domain (5 open ended questions and 7 items spread between the libido and sexual function subdomains), libido subdomain (3 items), sexual function subdomain (4 items), energy domain (3 items), sleep domain (3 items), cognition domain (3 items), and mood domain (7 items). Subjects will complete the HIS-Q as indicated in Appendix C[1]. The example HIS-Q is provided in Appendix H in the Protocol[1].

- **Patient Global Index of Improvement (PGI – I)**
  The PGI-I is adapted from questions that assess two other conditions, severity and improvement questionnaires in the treatment of men with lower urinary tract symptoms secondary to BPH[5] and global impression questionnaires for incontinence.[6] Subjects not participating in either the sexual function or the PDD sub-studies will be asked about their global impression of hypogonadism overall, while men participating in the sexual function sub-study and/or PDD sub-study will be asked about their impression of their libido and/or depression, respectively, since starting study drug. The PGI-I will be completed as indicated in Appendix C (for all subjects in main study) and Appendix O (for subjects who qualify for sexual function and/or PDD sub-studies only). The example PGI-I questions are provided in Appendix I in the Protocol[1].

- **International Index of Erectile Function (IIEF-5) – Sexual Function Sub-Study Only**

The IIEF-5 is a 5-item self-administered questionnaire measurement of erectile dysfunction.[7] Subjects will complete the IIEF-5 questions as indicated in Appendix O. The example IIEF-5 questions are provided in Appendix J[1].

- **Psychosexual Daily Questionnaire (PDQ) Question 4 Only – Sexual Function Sub-Study Only**
  The PDQ[8] is a self-reporting instrument designed for the assessment of sexual function and mood on a daily basis. The subjects will be asked to complete Question 4 from the PDQ for 7 consecutive days before each study visit. Subjects will complete the PDQ Question 4 as indicated in Appendix O. The example PDQ Question 4 is provided in Appendix K in the Protocol[1].

### 2.3.2 Primary Endpoint

- Difference between TRT and Placebo in change from baseline in overall sexual activity score, over a 1-year study period, using Question 4 of the PDQ.

### 2.3.3 Secondary Endpoints

- Difference between TRT and Placebo in change from baseline in sexual desire score, over a 1-year study period, using libido domain of HIS-Q instrument.

- Difference between TRT and Placebo in change from baseline in hypogonadal symptoms over a 1-year study period, assessed using the HIS-Q instrument.

  Difference between TRT and Placebo in change from baseline in sexual symptoms of hypogonadism over a 1-year study period, asssessed using the sexual domain of the HIS-Q instrument.
- Difference between TRT and Placebo in change from baseline in erectile function score, over 1-year study period, using erectile function domain of the IIEF.

- Change in the difference between TRT and Placebo in change from baseline in sexual activity score, sexual desire and erectile function at 1-year vs. 2-year.

- Relative risk for TRT vs. Placebo of clinically meaningful increase from baseline in PDQ Question 4 score (≥ 0.6) at month 6, 12 and 24.

- Difference between TRT and placebo in change from baseline in the libido domain of the HIS-Q questionnaire over the course of the 3-year follow-up.

- Difference between TRT and Placebo in change from baseline in hypogonadal symptoms assessed using the HIS-Q instrument over the course of the 3-year follow-up

- Determine whether, compared to placebo, TRT improves sexual symptoms of hypogonadism assessed using the HIS-Q instrument over the course of the 3-year follow-up

- The distribution of hypogonadism symptoms (study inclusion criteria) at baseline among all participants enrolled in the parent trial

### 2.3.4 Exploratory Endpoints

- Difference between TRT and placebo in change from baseline in the sexual function domain of the HIS-Q in all participants enrolled in the parent trial
- Difference between TRT and placebo in change from baseline in the hypogonadal symptoms of the HIS-Q instrument in all participants enrolled in the parent trial
- Difference between TRT and placebo in change from baseline in the sexual symptoms of hypogonadism assessed using the HIS-Q instrument in all participants enrolled in the parent trial
- Difference between TRT and placebo in sexual function using the patient global impression of change (PGI-I Libido) question.
- Evaluation of minimally important difference (MID) for the score obtained from HIS-Q libido question.

- Evaluation of the association between changes in testosterone levels and changes in sexual function outcomes at year 1, in patients randomized to testosterone arm.

## 2.4 Sample Size Justification

We anticipate data from at least 810 participants will be included in the Sexual Function Sub-study. This sample will provide more than 90% power to detect standardized testosterone effect as low as 0.3 between treatment and placebo, assuming 15% missing records at 1 year and using a two-sided two-sample T-test with significance alpha level of 0.05.
We have assumed a much smaller treatment effect than that which was observed in other testosterone trials[9-11] because the treatment effect tends to wane over time such that the average treatment effect may be less than that observed in other short-term testosterone trials. Also, because of the substantially larger size and longer duration of this trial relative to previous trials and because the trial will be conducted at approximately 400 trial sites, the rate of missing questionnaire data might be higher (15% annually) than in other trials, including the T Trial.[9]

The treatment effects for sexual function outcomes in the T Trials, and other testosterone trials, were larger than the effects assumed here. Therefore, a sample size of 810 men will provide robust statistical power to detect clinically meaningful treatment effects in sexual function scores.

## 2.5 Interim Analysis

No interim analysis is planned for this sub-study.

## 2.6 Multiplicity Testing Procedures for Type-I Error Control

Type I error adjustments for multiple comparisons are not planned for efficacy endpoints, subgroup analyses, supportive analyses or sensitivity analyses for this sub-study.

### 2.7 Missing Data

We have no intention to impute data in this sub-study.

If required (e.g. by referees in publication), multiple imputation of endpoints and covariates by the MICE methodology[12,13] for the report of sexual function sub-study results may be considered where appropriate (i.e. for data reported in the manuscripts). This method is notable for being able to handle clustering of repeated measures at the participant level, a feature of the design of this trial and sub-study. Under such circumstances we would consider whether to impute data for all sub-studies simultaneously.

## 3.0 Analysis Populations and Important Subgroups

### 3.1 Analysis Population

The Efficacy Analysis Set pre-specified for Sexual Function sub-study will be a subset of the FAS (Full Analysis Set) comprising all randomized subjects in the main CV study analyzed according to treatment assigned at randomization. Eligible subjects from the main study who satisfy the criteria for sub-study for sexual function will be included in the analysis. The EAS for the analysis of sexual function sub-study will include subjects with Baseline DISF-SRII Section I: Sexual Desire domain score $\leq 20$.

Subjects will be categorized according to treatment assigned at randomization. The EAS will be mainly used for the summary of subjects' disposition and summary of subjects' demographics and baseline characteristics for the sub-study.

Statistical analyses of treatment effect over time will comprise all subjects from the sub-study, who have baseline and at least one post-randomization measurement.

A modified analysis set for participants censored at any time (and subsequently) that they are deemed treatment noncompliant per the parent trial protocol[1], will be also considered and used for sensitivity analyses of efficacy endpoints.

Data from eligible measurements will be included, where eligible is defined as being obtained from questionnaires or scores for which at least 80% of items are non-missing.

### 3.2 Subgroup analyses

The following subgroup analysis will be carried out for the main study and may be also considered for sexual function efficacy sub-study:

- by race
- by age (< 65 year, ≥ 65 years)
- by baseline antidepressant use
- by prior CV disease status (yes, no)
- by baseline total testosterone levels (< 250 mg/dl, ≥ 250 mg/dl)

Additional sub-group analyses for sexual function sub-study, i.e. by age groups (other than 65 years), type II diabetes status or by baseline prostate symptom score (I-PSS) may be also carried out as appropriate.

## 4.0 Analysis Conventions

**The Operating Procedures of Data Collection and Analysis.** In case the following data operating procedures do not cover any particular aspect of data analysis or variable definitions in this sub-study Statistical Analysis Plan, the operating procedures of the Parent TRAVERSE Trial will be binding. For participants with duplicate ID in the database, data from the duplicated subjects will be excluded from the corresponding analysis sets in this sub-study. An average number of days in a year will be set as 365.25, and an average number of days in 1 month will be set as 365.25/12 for exposure calculation and time to

event analyses. The pre-existing cardiovascular disease will be defined, in accordance with parent trial definition, as an occurrence of one or more of the three conditions: coronary artery disease, cerebrovascular disease or peripheral arterial disease. If the same questionnaire data is collected more than once on the same day, the worst score (the highest or the lowest one, depending on the questionnaire) will be used from that day. If there are multiple assessments around the nominal day, the assessment closest to the nominal day will be used.

### 4.1 Definition of Baseline

Baseline on each outcome measure will be defined as the last available measurement obtained prior to the first dose of study drug (defined as on or before Day 1) (Protocol Appendix C).

### 4.2 Definition of Final Observation

Final observation for each analysis will be the final assessment affiliated with the relevant visit. For example, for twelve-month assessment of PDQ instrument, the final observation of each relevant measurement affiliated with the 12-month visit will be included in analyses.

### 4.3 Definition of Visit Windows

Definitions of the visit windows (baseline and on-treatment) are presented in Section 10.0 (Tables 10.1 and 10.2) and schedule of sub-study activities is presented in Section 9.0.

## 5.0 Demographics, Baseline Characteristics, Medical History and Study Drug Exposure

### 5.1 Baseline Characteristics

Data collected in this sub-study will be documented using summary tables. Statistics for continuous variables will include mean, median, standard deviation, minimum, maximum,

and sample size for each treatment group, and two-sided 95% confidence intervals of the mean difference between the treatment groups. Binary variables will be described with frequencies, percentages, and two-sided 95% confidence intervals of the difference in percentages between treatments.

Medical history will mirror the presentation in the parent TRAVERSE trial, and will include at minimum history of depression, CV history; nicotine and alcohol use; and testosterone use.

### 5.2 Report of Treatment Exposure and Compliance

Continuous summaries of subjects' total duration of treatment with study drug, among participants recruited in Sexual Function sub-study, will be provided for analyzed time intervals (1- and 2-year follow-up).
Total patient-month of exposure will be calculated by summing the duration of treatment (separately for 1- and 2-year follow-up) for all subjects in the analysis set and dividing this sum by 365.25 (= 1 year). In addition, the number and percentage of subjects exposed to study drug will be summarized for the following categories of exposure duration: ≤ 1 month, > 1 to 2 months, > 2 to 3 months, > 3 to 6 months, > 6 months to 1 year, > 1 to 2 years, etc.
Study drug compliance will be computed for sexual function sub-study participants, separately for 1 and 2-year study intervals.

## 6.0 Analysis of Endpoints

### 6.1 Primary Analysis

Linear mixed model will be employed to compare effect of 1-year intervention on the change over time in the sexual activity (score from Question 4 PDQ) against placebo. Change will be defined as the difference from the baseline value. Models will be adjusted according to whether participants had a prior CV event (a stratifying factor in the parent trial) and will consist of visit, treatment effect, visit-by-treatment interaction and baseline value as fixed effects. Random intercept will be included at participants' level. Unstructured covariance matrix will be assumed, however if convergence of the model is not achieved, then a compound symmetry structure will be utilized. Effects of overall sexual activity

change over 1-year treatment period will be calculated as an average score from 6 and 12-month visits, and will be extracted, along with two-sided 95% confidence intervals, from the mixed-model framework.

Sensitivity analyses of treatment effect may also be performed where appropriate (e.g. inclusion in the model other fixed effects such as stratification factors etc., or including only those individuals compliant and completing the intervention period, as noted above).

## 6.2 Analysis of Secondary Endpoints

Effect of 1-year treatment intervention on sexual desire score (libido domain of HIS-Q instrument), hypogonadal symptoms and sexual function (HIS-Q sub-domains) will be examined in similar fashion as primary endpoint analyses. Linear mixed model will be employed to compare treatment effect over 1-year follow-up. Change in sexual desire score will be defined as the difference from the baseline value. Random intercept model will consist of visit, treatment effect, visit-by-treatment interaction and baseline value as fixed effects and will be adjusted to prior CV event. All models will assume unstructured covariance matrix (or compound symmetry structure if model will not converge). Effects of sexual desire score change over 1-year treatment period will be expressed as an average score from 6 and 12-month visits, and will be extracted, along with two-sided 95% confidence intervals, from the mixed-model framework.

Effect of 1-year treatment intervention for erectile function will be performed using analysis of covariance model (ANCOVA). Outcome will be expressed as change from Baseline at year 1 in erectile function domain score of IIEF-5 instrument. Model will be adjusted to prior CV event and will consist baseline value and treatment group as the independent variables. Effects of 1-year intervention for erectile function score change from baseline will be expressed as the difference between placebo and treatment arms at 12-month visits, and will be extracted, along with two-sided 95% confidence intervals, from the ANCOVA model.

To investigate if improvements in sexual activity, sexual desire, and erectile function over 1 year are maintained after 2 years of treatment versus placebo we will employ mixed-model regressions (overall sexual activity score Question 4 from PDQ, domains from

HIS-Q and IIEF-5) with inclusion of data for all available time points over 2-year follow-up. Change in hypogonadal symptoms, sexual activity, desire and erectile function will be defined as the difference from the baseline value. Random intercept model adjusted to prior CV event will consist of visit, treatment effect, visit-by-treatment interaction and baseline value as fixed effects. All models will assume unstructured covariance matrix (or compound symmetry structure if model will not converge). To investigate continuation of treatment effect vs. placebo after 12 months of intervention, estimates and 95% two-sided confidence intervals for 1-year and 2-year visits will be extracted from mixed-model framework and the difference between those two timepoints will be tested with use of treatment contrasts (Contrast Statement in SAS Proc Mixed[14]).
To examine continuation of treatment effect at year 2 for PGI-I Libido domain score we will perform analysis of covariance model. Outcome will be expressed as change between 2- and 1-year score and model will be adjusted to year 1 value and prior CV occurrence. Difference between those two timepoints will be tested with use of treatment contrasts.

Evaluation of TRT effect on clinically meaningful improvement in PDQ Question 4 score will be analyzed using Chi-square test.[15,16] Outcome will be defined as an increase in the score by 0.6 or more units[17] between baseline and 6-, 12- and 24-month visit and thus calculated frequencies will be compared between treatment arms.. Evaluation of TRT effect on clinically meaningful improvement in other sexual activity metrics might also be considered and these analyses will be conducted in similar fashion.

Evaluation of long-term testosterone effect on the libido domain of the HIS-Q questionnaire at the end of follow-up (year 5) will be performed using ANCOVA model and these analyses will be executed in similar manner as other secondary endpoints analyses.

The distribution of participants with one ore more hypogonadism symptoms at baseline: decreased sexual desire or libido, decreased spontaneous erections, decreased energy or fatigue/feeling tired, low mood or depressed mood, loss of body axillary and pubic hair or reduced shaving or hot flashes) among all men enrolled in the parent trial will be assessed numerically (numbers and frequencies) and graphically.

Sensitivity analyses of treatment effect may also be performed where appropriate (i.e. inclusion in the model other stratification factors etc.).

## 6.3 Analysis of Exploratory Endpoints

The effect of treatment versus placebo therapy on patient global impression of change in sexual function will be examine using a patient global impression of change (PGI-I Libido) question. ANCOVA model will be employed to compare difference between groups in at 1-year visit. Model will include treatment effect and will be adjusted to prior CV event. Treatment effect will be expressed as a score at 12-month visits, and will be extracted, along with two-sided 95% confidence intervals, from the ANCOVA model.

Sensitivity analyses of treatment effect may also be performed where appropriate (e.g. inclusion in the model other fixed effects such as stratification factors etc.).
Derivation of minimally important difference (MID) will be performed using anchor-based methods for patients on testosterone treatment.[18] In this approach global impression of change (PGI-I Libido) will be used as an anchor and the change in score for HIS-Q libido question will be used for evaluation of MID. Furthermore, other methods might be considered to derive aggregate results across approaches, specifically distribution-based method, where MID threshold is derived with respect to 1/2 and 1/3 standard deviation change in HIS-Q score.[19]

Exploratory graphical assessment of the relationship between change in testosterone levels and change in sexual function outcomes at year 1 will be performed for all primary and secondary outcomes. Where appropriate, log or power transformations of variables will be employed to enhance conformity with normality assumptions. Association between dependent and independent variables will be evaluated using linear regression models and potential non-linear relationship will be inspected using restricted cubic splines.[20-21]

## 6.4 Safety Endpoint

Safety assessments will be incorporated into the parent trial. No additional safety endpoints will be specified in this sub-study.

# 7.0 Summary of Changes

## 7.1 Summary of Changes Between the Previous Version and the Current Version

Not applicable.

## 7.2 Summary of Changes in Previous Version

Not applicable

# 8.0 References

1. Study protocol of M16-100, Amendment 1. 26 February 2018.

2. Barry MJ, Fowler FJ Jr, O'Leary MP,et. al. The American Urological Association symptom index for benign prostatic hyperplasia. The measurement committee of the American urological association. Journal of Urology. 1992;148(5):1549-57.

3. Derogatis LR. The Derogatis Interview for Sexual Functioning (DISF/DISF-SR): an introductory report. Journal of Sex and Marital Therapy. 1997; 23(4):291-304.

4. Gelhorn HL, Vernon MK, Stewart KD et al. Content Validity of the Hypogonadism Impact of Symptoms questionnaire (HIS-Q): a patient-reported outcome measure to evaluate symptoms of hypogonadism. Patient. 2016;9(2):181-90.

5. Viktrup L, Hayes RP, Wang P, et al. Construct validation of patient global impression of severity (PGI-S) and improvement (PGI-I) questionnaires in the treatment of men with lower urinary tract symptoms secondary to benign prostatic hyperplasia. BMC Urol. 2012; 12:30.

6. Yalcin I, Bump RC. Validation of two global impression questionnaires for incontinence. Am J Obstet Gynecol. 2013;189(1):98-101.

7. Rosen RC, Riley A, Wagner G, et al. The international index of erectile function (IIEF): a multidimensional scale for assessment of erectile dysfunction. Urology.1997;49(6):822-30.

8. Lee KK, Berman N, Alexander GM, et al. A simple self-report diary for assessing psychosexual function in hypogonadal men. J Androl. 2003;24(5):688-98.

9. Snyder PJ, Bhasin S, Cunningham GR, et al. Effects of testosterone treatment in older men. N Engl J Med. 2016;374(7):611-24.

10. Brock G, Heiselman D, Maggi M, et al. Effect of Testosterone Solution 2% on Testosterone Concentration, Sex Drive and Energy in Hypogonadal Men: Results of a Placebo Controlled Study. J Urol. 2016;195(3):699-705.

11. Cunningham GR, Stephens-Shields AJ, Rosen RC, et al. Testosterone Treatment and Sexual Function in Older Men with Low Testosterone Levels. J Clin Endocrinol Metab. 2016;101(8):3096-104.

12. Van Buuren, S., Flexible Imputation of Missing Data. 2012, Chapman & Hall/CRC.

13. Van Buuren, S. and K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 2011. 45(3): p. 67.

14. SAS Institute Inc. SAS/STAT® 13.1 User's Guide. Cary, NC: SAS Institute Inc.; 2013.

15. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. American J Epidemiol 2004; 159(7): 702–706.

16. Rivers, C., M.S. Majumder and E.T. Lofgren. Risks of Death and Severe Disease in Patients with Middle East Respiratory Syndrome Coronavirus, 2012–2015. Am J Epidemiol. 2016;184(6): 460-4.

17. Wang C, Stphens-Shields AJ, DeRogatis LR, et al. Validity and Clinically Meaningful Changes in the Psychosexual Daily Questionnaire and Derogatis Interview for Sexual Function Assessment: Results from the TTrials. J Sex Med. 2018;15(7):997-1009.

18. Copay AG, Subach BR, Glassman SD, Polly DW, Jr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 2007;7: 541–546.

19. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *J Clin Epidemiol* 2004,57(9):898–910.

20. Harrell, FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag 2010, New York, Inc. New York, USA.

21. Steyerberg, WW. Clinical Prediction Models. Springer-Verlag 2009, New York, Inc. New York, USA.

# 9.0 Partial List of Tables with Schedule of Activities

## 9.1 Study Activities (Sexual Function Questionnaires)

| Activity | SV1 | SV2 | SV3 | D1 (Baseline) | W 2 | M 1 W 4 | M 3 W 12 | M 6 W 26 | M 9 W 39 (phone) | M 12 W 52 (year 1) | M 15 W 65(phone) | M 18 W 78 | M 21 W 90 (phone) | M 24 W 104 (year 2) | M 27 W 116 (phone) | M 30 W 130 | M 33 W 142 (phone) | M 36 W 156 (year 3) | M 39 W 168 (phone) | M 42 W 182 | M 45 W 194 (phone) | M 48 W 208 (year 4) | M 51 W 220 (phone) | M 54 W 234 | M 57 W 246 (phone) | M 60 W 260/FV (Year 5) | PD | Unscheduled | 30-Day Call |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PRO** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| I-PSS[d] | | X | | X | | | X | | | X | | | | | | | | X | | | | | | | | X | X | | |
| DISF – SRII Section I[e] | | X | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| HIS-Q | | | | X | | | | X | | X | | | | X | | | | | | | | | | | | X | X | | |
| PHQ-9 | | X | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PGI-I Hypogonadism | | | | | | | | X | | | | | | X | | | | | | | | | | | | X | X | | |

d. I-PSS greater than 19 is ONLY exclusionary at SV2.

e. Subjects with a DISF – SRII Section I ≤ 20 at SV2 (Appendix G in the Protocol[1]) and consented to participation in the sexual function sub-study (Appendix O) need to complete the 7-day PDQ Question 4 (Appendix K) prior to Baseline (Day 1) Visit.

## 9.2 Additional Sexual Function Sub-Study Activities

| Activity | 7 Consecutive Days Prior to Day 1 | D1 | M 6 W 26 | M 12 W 52 | M 18 W 78 | M 24 W 104 | M 60 W 260/FV (Year 5) | PD |
|---|---|---|---|---|---|---|---|---|
| IIEF-5 | | X | | X | | X | | |
| PDQ Question 4[a] | X | | X | X | | X | | |
| PGI-I Libido | | | | X | | X | X | X |

a. Subjects meeting the inclusion criteria for sexual function sub-study following SV2 (above) need to complete the 7-day PDQ Question 4 (Appendix K in the Protocol[1]) prior to Day 1 Visit (Baseline).

# 10.0 Efficacy Analysis Time Windows

## 10.1 Visit windows for PDQ-4, HIS-Q and PGI- I libido (no baseline)

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1[a] | 1 | ≤ 1 |
| M6 | 182 | 92 - 273 |
| M12 | 364 | 274 - 546 |
| M24 | 728 | 547 - 1093 |
| M60[b] | 1820 | 1460 – 2180 |
| Final Visit | | 2 to ≤ 2 days after the last dose of study drug |

a. Only for PDQ-4 and HIS-Q
b. Only for HIS-Q and PGI-I Libido

### 10.2 Visit windows for IIEF-5

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M12 | 364 | 274 - 546 |
| M24 | 728 | 547 - 1093 |
| Final Visit | | 2 to ≤ 2 days after the last dose of study drug |

# 11.0 Examples of Table Shells and Figures

## 11.1 Sexual Function Sub-Study Outcomes

| **Outcome** | No. of Men | Baseline Value | Change from Baseline Value[1] | | Treatment Effect (95% CI) | P Value |
|---|---|---|---|---|---|---|
| | | | Month 6 | Month 12 | | |
| Primary outcome: PDQ-Q4 score | | | | | | |
| Testosterone | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx.x (xx.x-xx x) | x xxx |
| Placebo | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx.x (xx.x-xx x) | x xxx |
| Secondary outcomes | | | | | | |
| HIS-Q Libido score | | | | | | |
| Testosterone | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx.x (xx.x-xx x) | x xxx |
| Placebo | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx.x (xx.x-xx x) | x xxx |
| IIEF-5 Erectile Function score | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Testosterone | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx.x (xx.x-xx x) | x xxx |
| Placebo | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx.x (xx.x-xx x) | x xxx |

[1]Model-based estimates and 95% CIs at each time point.

## 11.2 Comparison of TRT Effect between 1 and 2-year Study Intervention

| **Outcome** | No. of Men | Baseline Value | Change from Baseline Value[1] | | Treatment Effect[2] (95% CI) | P Value |
|---|---|---|---|---|---|---|
| | | | Month 12 | Month 24 | | |
| Primary outcome: PDQ-Q4 score | | | | | | |
| Testosterone | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx x) | x xxx |
| Placebo | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx x) | x xxx |
| Secondary outcomes | | | | | | |
| HIS-Q Libido score | | | | | | |
| Testosterone | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx x) | x xxx |
| Placebo | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx x) | x xxx |
| IIEF-5 Erectile Function score | | | | | | |
| Testosterone | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx x) | x xxx |
| Placebo | N = | xx x ± xx x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx x) | x xxx |

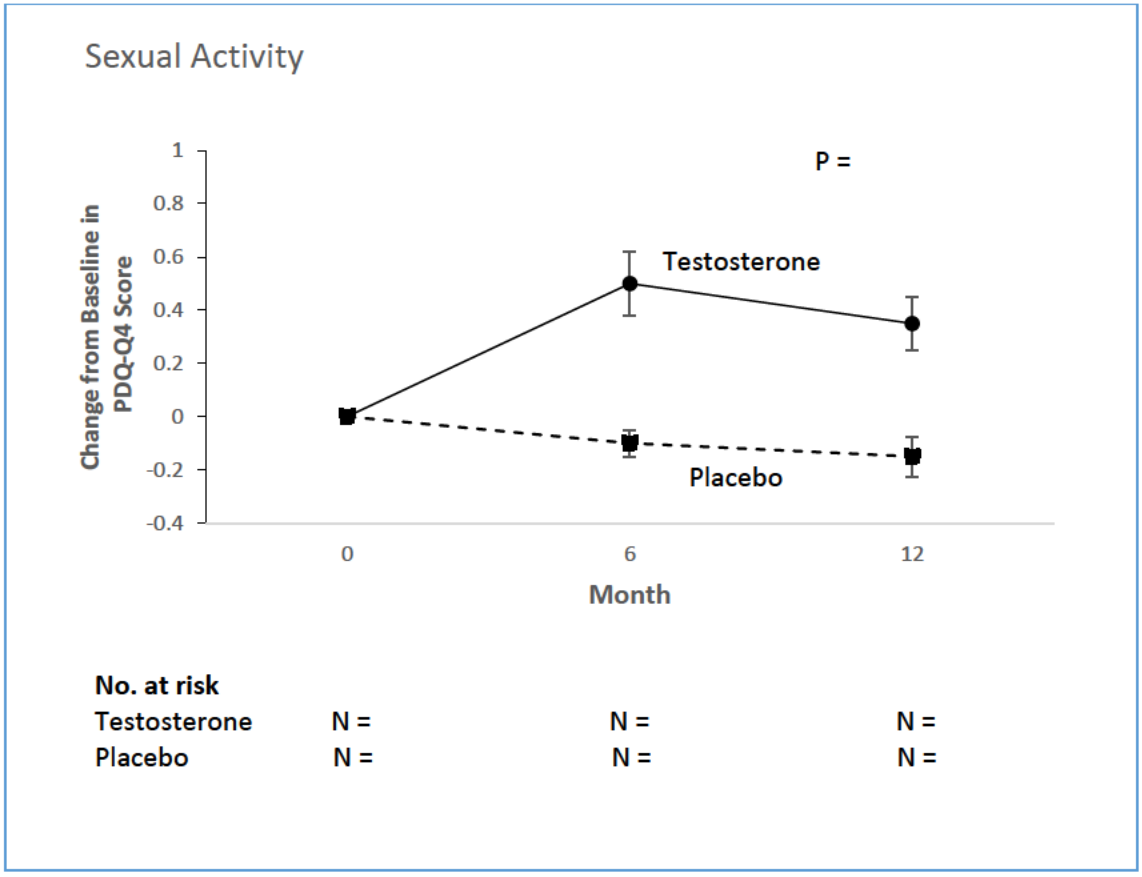
[1]Model-based estimates and 95% CIs at each time point. Treatment effect expressed as change between 1 and 2-year intervention.

## 11.3 Continuation of TRT Effect for PGI-I Libido Domain at 2-year versus 1-year Intervention

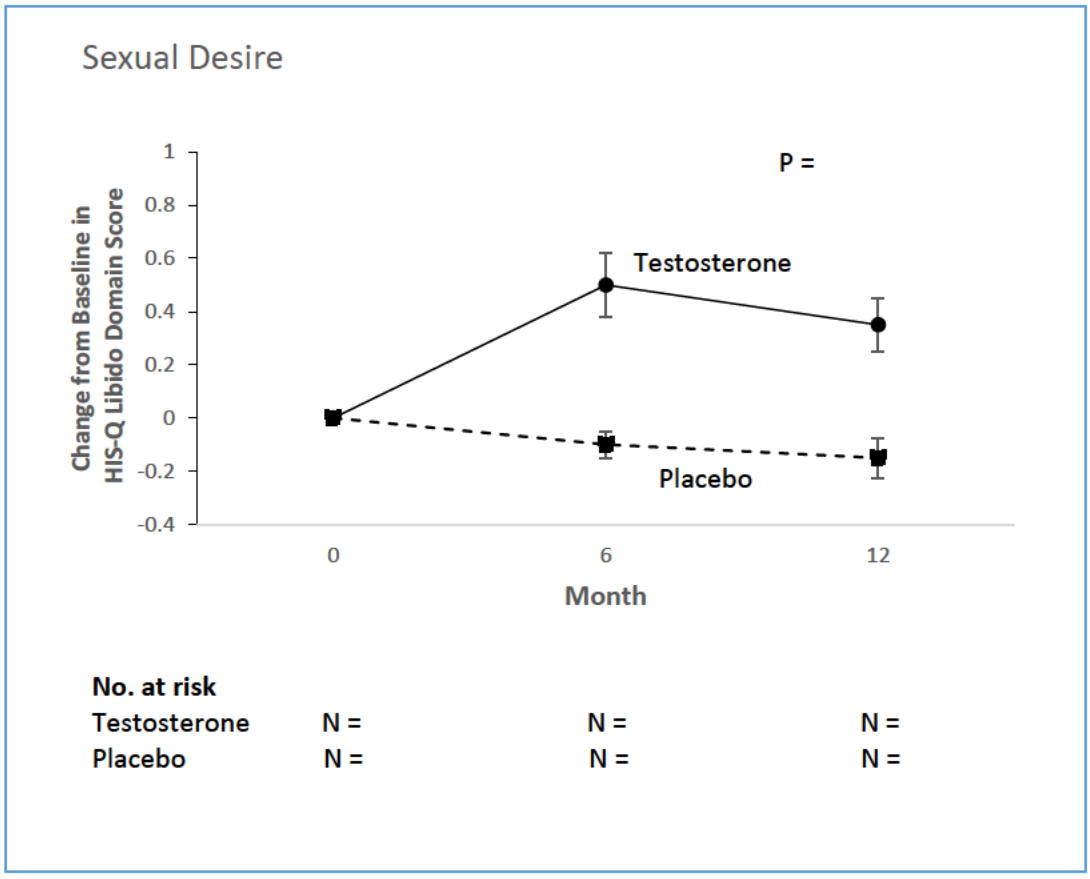
| **Outcome** | No. of Men | Baseline Value | Baseline Value[1] | | Treatment Effect[2] (95% CI) | P Value |
|---|---|---|---|---|---|---|
| | | | Month 12 | Month 24 | | |
| Exploratory outcome: PGI-I Libido Domain score | | | | | | |
| Testosterone | N = | xx x ± xx.x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx.x) | x xxx |
| Placebo | N = | xx x ± xx.x | xx x ± xx x | xx x ± xx x | xx x (xx x-xx.x) | x xxx |

[1]Model-based estimates and 95% CIs at each time point. Treatment effect expressed as change between 1 and 2-year intervention.

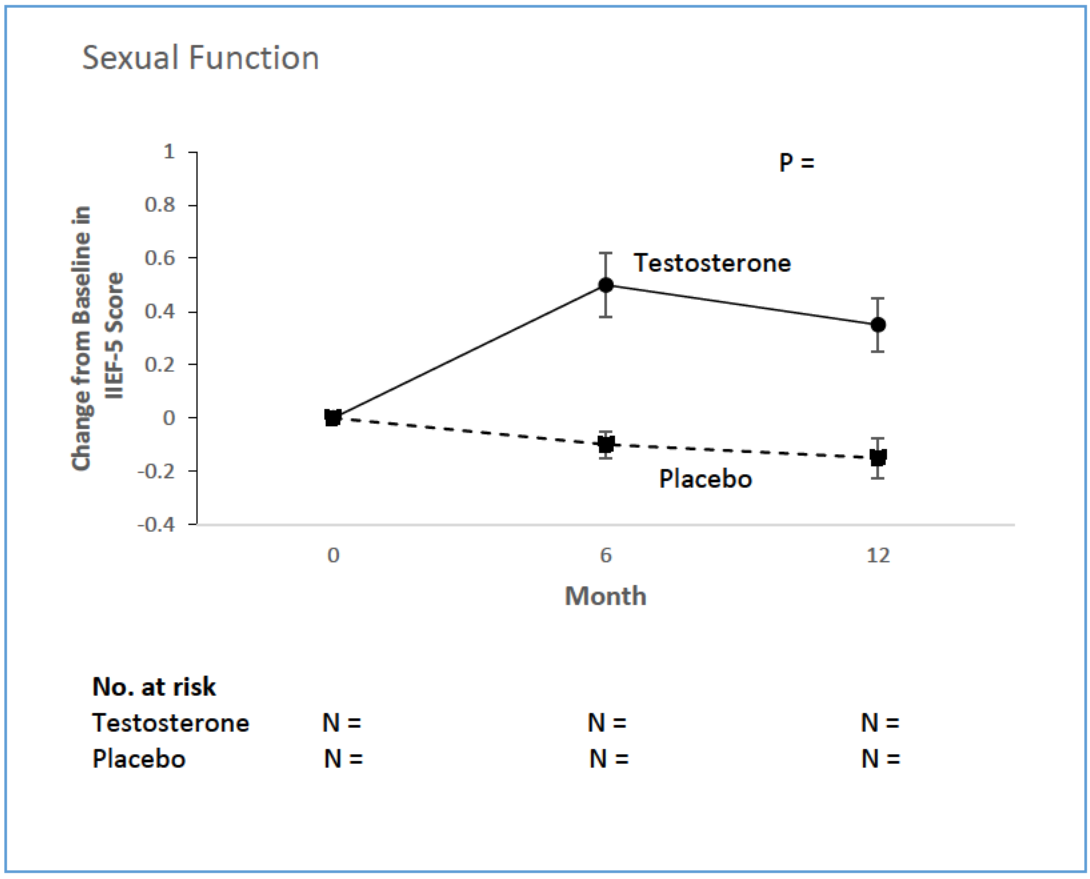
## 11.4 Sexual Activity Score (PDQ Question 4) Change from Baseline over 1-year Follow-up



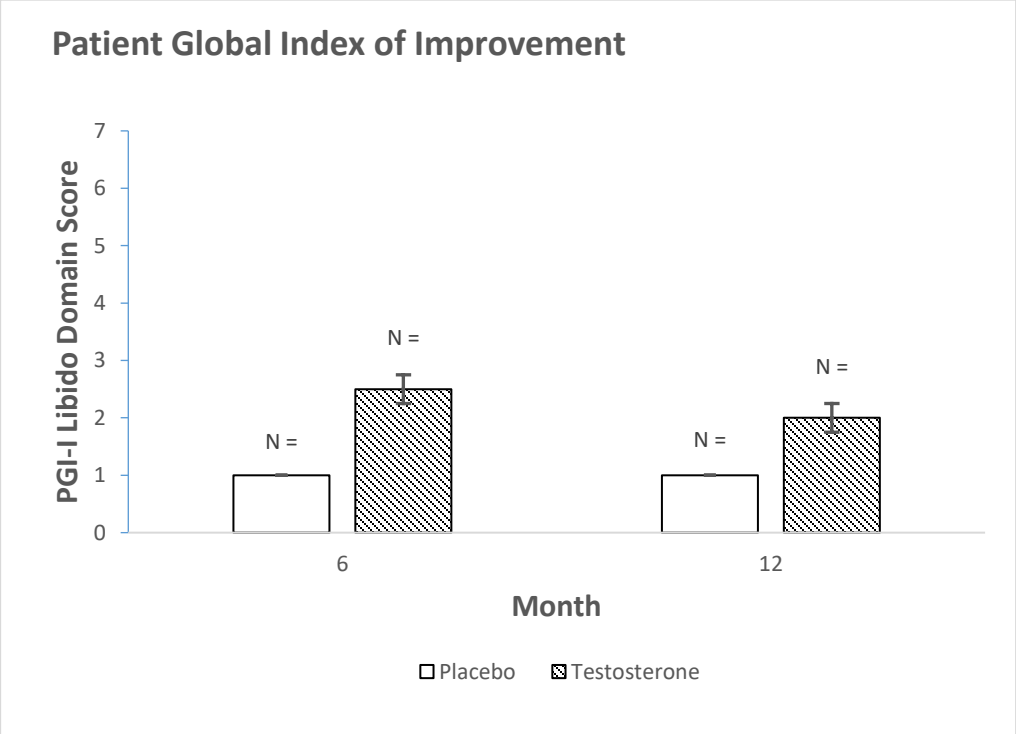
## 11.5 Sexual Desire Score (HIS-Q Libido domain) Change from Baseline over 1-year Follow-up



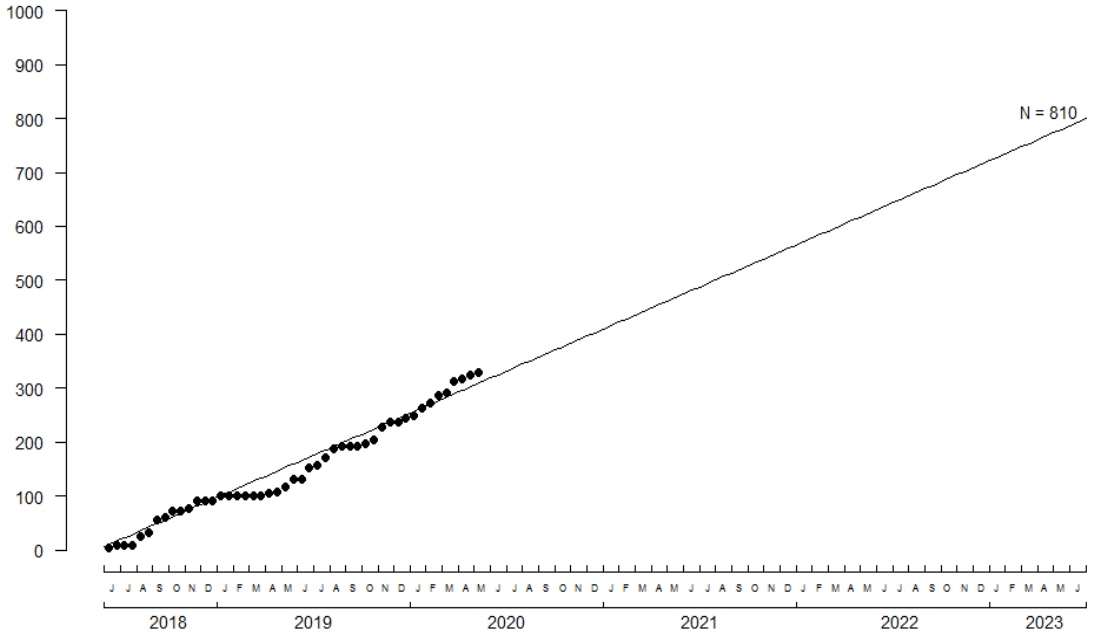
## 11.6 Sexual Function Score (IIEF-5) Change from Baseline over 1-year Follow-up



## 11.7 Patient Global Change of Impression in Sexual Function (HIS-Q Libido domain) at 1- and 2-year Follow-up



## 11.8 Enrollment Progress for Sexual Function Efficacy Sub-Study



# Statistical Analysis Plan

## Supplement for Analyses of Prostate Safety Endpoints

**Study M16-100**
**Testosterone Replacement Therapy for Assessment of Long-Term Vascular Events and Efficacy ResponSE in Hypogonadal Men (TRAVERSE) Study**

**Date: January 01, 2023**

**Table of Contents**

**List of Abbreviations**

| | |
|---|---|
| 5-ARI | 5-Alpha Reductase Inhibitor |
| AUR | Acute Urinary Retention |
| CV | Cardiovascular |
| CVD | Cardiovascular Disease |
| DVD | Digital Versatile Disc |
| ESC | Executive Steering Committee |
| I-PSS | International Prostate Symptom Score |
| IRB | Institutional Review Board |
| IRT | Interactive Response Technology |
| MACE | Major Adverse Cardiac Event |
| MICE | Multiple Imputation by Chained Equations |
| PCPT | Prostate Cancer Prevention Trial |
| PDQ | Psychosexual Daily Questionnaire |
| PSA | Prostate Specific Antigen |
| SAP | Statistical Analysis Plan |
| TRT | Testosterone Replacement Therapy |
| TURP | Transurethral Resection of the Prostate |

## 1. Introduction

This statistical analysis plan (SAP) provides details to elaborate statistical methods for the analyses of prostate safety endpoints collected as outlined in the protocol of Study M16-100 parent trial (Amendment 1 dated 26 February 2018[1]) and to describe analysis conventions to guide the statistical programming work.

All analyses will be performed using SAS Version 9.2 or higher (SAS Institute, Inc., Cary, NC 27513) or R Version 3.4 or higher. This SAP will not be updated in case of future (administrative or minor) amendments to the TRAVERSE trial protocol unless the changes have any impact on the analysis of study data described here.

Analytic approaches described below are presented under the expectation that there will be sufficient information to render them meaningful. However, it is possible that the number of prostate safety events (e.g. new cancers and/or biopsies) may not be large enough to provide sufficient statistical power for some suggested analyses. In this case, only descriptive analyses may be performed.

## 2. Study Background

## 3. Objective

**Primary Objective:**

- To compare the incidence of high-grade prostate cancer (Gleason score of 4 +3 or higher) in middle-aged and older hypogonadal men randomized to TRT or placebo gel.

**Secondary Objectives:**

- To compare the frequency of prostate biopsies in men treated with TRT and placebo gel.
- To compare the incidence of any prostate cancer in hypogonadal men treated with TRT and placebo gel.
- To compare changes in PSA levels over time in men treated with TRT and placebo.
- To compare the incidence of acute urinary retention in men treated with TRT and placebo gel.
- To compare the percentage of men starting pharmacologic treatment for lower urinary tract symptoms in the TRT and placebo groups.
- To compare the percent of men undergoing invasive prostate surgical procedures (e.g., prostatectomy, transurethral prostate resection, or other prostate surgical procedure) for benign prostatic hyperplasia (BPH) in the TRT and placebo groups.
- To compare the effects of TRT and placebo on change in I-PSS score over the entire study period.

- To compare the percent of men undergoing treatment for prostate cancer (radical prostatectomy, radiation therapy, focal ablative therapy) in the TRT and placebo groups.
- Analyses to assess bias in ascertainment of prostate events

## 4. Study Design

The TRAVERSE parent trial is a Phase 4, randomized, double-blind, placebo-controlled, multicenter study of topical TRT in symptomatic hypogonadal men with increased risk for CV disease. The initial planned study enrollment is approximately 6,000 subjects based on the projected timing when 256 MACE will occur under initial assumptions of the annual event rate, subject accrual rate, and study discontinuation rate. There will be approximately 400 sites in North America. An Interactive Response Technology (IRT) system will randomize subjects to receive either topical testosterone or placebo in a 1:1 ratio. Randomization will be stratified by pre-existing CV disease (Yes/No). Titration of testosterone dose will occur in subjects receiving active testosterone, while sham dosage titrations will occur in subjects receiving placebo gel via the central IRT system. The Screening Period is up to 60 days prior to first dose of study drug. Once subjects meet all of the eligibility criteria during Screening, they will be randomized (1:1 ratio) to active study drug or placebo and will be followed until the study ends, with stratification by prior CVD event status (see below). Importantly, randomized subjects who elect to discontinue study drug will also be followed until the study ends unless the subject dies or withdraws from the study completely (withdrawal of informed consent) earlier. Subjects who discontinue study drug will still be asked to follow their regularly scheduled protocol visits. Subjects who interrupt study drug will be allowed to restart study drug at any time.

### 4.1. Background

Evidence of a relationship between TRT and the incidence of prostate cancer is mixed. In the Baltimore Longitudinal Study of Aging, aggressive prostate cancers were reported to be associated with higher levels of total and free testosterone.[1] Also, testosterone administration increases prostate specific antigen (PSA) in hypogonadal men and can promote the growth of metastatic prostate cancer.[3] On the other hand, most population-based studies have not associated high total or free testosterone levels with increased cancer risk,[4-6] and an analysis of the placebo arm from the Prostate Cancer Prevention Trial (PCPT) found no significant associations of total or free testosterone and risk of total, low (Gleason < 7) or high-grade (Gleason 7 – 10) prostate cancer.[7,8] Occult prostate cancer is common in middle-aged and older men, and the prevalence increases with increases in PSA and age.[7,9] The designers of the current study recognize that testosterone therapy usually increases PSA in circulation and thus, in the current study it is likely that more men randomized to testosterone will be referred for prostate evaluation and possible biopsy based on this laboratory finding. Consequently, testosterone-treated men may have an increased risk of detection of subclinical prostate cancer that was present prior to treatment.[5,10,11] This inherent surveillance bias could result in a greater number of prostate biopsies that are positive for low-grade indolent prostate cancers in men randomized

to the testosterone arm than in those randomized to the placebo arm. With this recognition of the inherent detection bias from testosterone treatment leading to the detection of low grade prostate cancers, this study will assess the effect of testosterone and placebo on the development of high grade prostate cancer, defined as Gleason 4 + 3 or greater, as these higher grades are associated with increased morbidity and mortality. **All analyses will consider the potential for bias in ascertainment of events as part of sensitivity assessments.**

### 4.2. Variables Used for Stratification at Randomization

TRAVERSE trial randomization will be stratified by pre-existing CV disease (Yes/No). It is expected that in the parent trial 30% of the randomized subjects will satisfy inclusion criteria for pre-existing CV disease criteria (secondary prevention), and the remaining 70% will satisfy CV risk factors criteria (primary prevention) combined.

The ESC and Sponsor may decide to cap the primary prevention cohort if that cohort is found to consistently exceed 70% of the total population enrolled or if the pooled primary event rate falls below projections.

## 5. Endpoints

### 5.1. Primary Endpoint

- Time to first diagnosis of high-grade prostate cancer (Gleason score of 4 +3 or higher) defined as time from randomization to an event

### 5.2. Secondary Endpoints

- Incidence of high-grade prostate cancer (Gleason score of 4 +3 or higher)
- Incidence of any prostate cancer
- Time to prostate biopsies defined as time from randomization to an occurrence of prostate biopsy
- Comparison of cumulative incidence of one or more biopsies by 3, 12, and 24 months post-randomization
- Time to occurrence of first prostate cancer of any grade
- Time to occurrence of first acute urinary retention defined as time from randomization to an event
- Time to the first occurrence of starting pharmacologic treatment for lower urinary tract symptoms
- Time to the first occurrence of invasive prostate surgical procedures (e.g., prostatectomy, transurethral prostate resection, or other prostate surgical procedure) for benign prostatic hyperplasia

- Time to the first occurrence of composite endpoint: any grade prostate cancer, acute urinary retention, start of pharmacologic treatment for lower urinary tract symptoms, prostate biopsy or invasive prostate surgical procedures.

- Change from baseline in PSA levels and I-PSS score with time
- Change from baseline in Total and Free Testosterone, DHT, Estradiol and SHBG levels
- Gleason score and TNM stage of diagnosed prostate cancers
- Proportions of men with prostate biopsy, acute urinary retention, starting new pharmacologic therapy for lower urinary tract symptoms, and invasive prostate surgical procedures (e.g., prostatectomy, transurethral prostate resection,) for benign prostatic hyperplasia will be compared between the two intervention arms.
- Proportions of men undergoing treatment (radiation therapy, radical prostatectomy, focal ablation procedures) for prostate cancer will be compared between the two intervention arms.

## 6. Interim Analysis

No formal interim analysis or stopping rules are specified. Safety comparisons will be reported to the trial DSMB as stipulated in the parent TRAVERSE trial protocol.

## 7. Multiplicity Testing Procedures for Type-I Error Control

Type I error adjustments for multiple comparisons are not planned for this sub-study.

There is only one primary treatment comparison for the primary objective.

## 8. Missing Data Imputation

Following the TRAVERSE protocol, no imputation is planned. Multiple imputation of endpoints and covariates by the MICE methodology[12,13] may be considered where appropriate (i.e. for data reported in the manuscripts). This method is notable for being able to handle clustering of repeated measures at the participant level, a feature of the design of this trial.

## 9. Analysis Populations and Important Subgroups

### 9.1. Analysis population

The analyses will utilize the Safety Analysis Set comprising all subjects eligible for analysis in the main TRAVERSE trial, as defined in the study protocol[1]. Subjects will be categorized to the treatment arms according to the first treatment received, not the treatment assigned at

randomization. If a participant is assigned to Placebo at randomization but receives Androgel kit, the subject will be categorized to Androgel arm per safety analysis set.

A modified analysis set removing participants from the point at which they are deemed treatment noncompliant by the parent trial may also be considered and used in sensitivity analyses.

## 9.2. Analysis subpopulations

We will consider those individuals with biopsy in analyses of the rate of diagnoses per biopsy.

## 10. Pre-specified Subgroup Analyses

As noted above, analyses will be performed only where there are sufficient cases to support them. Under this stipulation, the following subgroup analyses will be considered for the prostate sub study endpoints:

- by race and ethnicity
- by age (e.g. < 65 year, ≥ 65 years)
- by family history of PCa (first degree relatives)
- by baseline PSA (dichotomized at median or other relevant value)
- by baseline total testosterone levels (e.g. < 250 mg/dl, ≥ 250 mg/dl)
- by baseline BPH (those with prevalent diagnoses vs those without)

Additional sub-group analyses may be also carried out as appropriate.

## 11. Analysis Conventions

**The Operating Procedures of Data Collection and Analysis.** In case the following data operating procedures do not cover any particular aspect of data analysis or variable definitions in this sub-study Statistical Analysis Plan, the operating procedures of the Parent TRAVERSE Trial will be binding. For participants with duplicate ID in the database, data from the duplicated subjects will be excluded from the corresponding analysis sets in this sub-study. An average number of days in a year will be set as 365.25, and an average number of days in 1 month will be set as 365.25/12 for exposure calculation and time to event analyses. The pre-existing cardiovascular disease will be defined, in accordance with parent trial definition, as an occurrence of one or more of the three conditions: coronary artery disease, cerebrovascular disease or peripheral arterial disease. If the same questionnaire data is collected more than once on the same day, the worst score (the highest or the lowest one, depending on the questionnaire) will be used from that day. If there are multiple assessments around the nominal day, the assessment closest to the nominal day will be used.

### 11.1. Definition of Baseline

Baseline on each outcome measure will be defined as the last available measurement obtained prior to the first dose of study drug (defined as on or before Day 1) (Protocol Appendix C).

### 11.2. Definition of Visit Windows

Definitions of the visit windows (baseline and on-treatment) are presented in Section 16 (Tables 16.1-16.3) and schedule of sub-study activities is presented in Section 15.

### 11.3. Referral to Prostate Biopsy Procedure

A subject will be referred for urological evaluation for consideration of further work-up which may include a prostate biopsy if he meets any of the following criteria:

- Confirmed increase > 1.4 ng/mL above Baseline during the first year [> 0.7 ng/mL in men on 5-Alpha Reductase Inhibitor (5-ARI)]
- Detection of a new prostate nodule or induration
- Confirmed absolute PSA value > 4.0 ng/mL at any time during the study (> 2.0 ng/mL in men on 5-ARI)

- Men 45 – 54 years of age whose Baseline PSA was < 1.5 ng/mL and whose PSA increases to > 3.0 ng/mL at any time during the study

For men aged 55 or older whose repeat PSA confirms the 1.4 ng/mL increase or a level > 4.0 ng/mL, their risk variables will be entered into the PCPT Risk Calculator Version 2.0.[14] The three resulting estimates will be provided: Risk of no cancer, risk of low-grade cancer, risk of high-grade cancer. These results will be calculated centrally and provided to the site and subject. With the subject's own risk estimates, he may then be provided with an IRB approved video (either a digital versatile disc (DVD) provided or an on-line video) that provides extensive and updated information about pros and cons of a prostate biopsy. Along with the video, the subject will be provided with a urology referral.

The IRT system will communicate the need for the site to refer subjects with confirmed increases in PSA to a urologist for further evaluation. Unblinded PSA results will be provided to the site once a repeat PSA is confirmed to meet the thresholds. It is recommended that study drug be continued in these subjects while being evaluated by a urologist. Subjects with non-confirmatory findings (e.g., negative biopsy, decision on the urologist's part not to do additional investigations) following the evaluation may continue on study drug. Subjects with biopsies positive for prostate cancer should have study drug discontinued but will continue all other trial procedures.

### 11.4. Adjudication of Prostate Endpoints

The following endpoints will be adjudicated by a Prostate Endpoints Adjudication Committee:

1. Prostate cancer status and Gleason score
2. Acute urinary retention events
3. Invasive prostate surgical procedure (prostatectomy, transurethral prostate resection, or other prostate surgical procedure) for obstruction

The diagnosis of prostate cancer is based on the evaluation of prostate biopsies and all prostate procedures that yield tissue and which are performed during the duration of the trial, including TURP and prostatectomy. Although great effort will be made to obtain materials to be reviewed by the TRAVERSE Prostate Adjudication Center at the University of Colorado, if the slides cannot be obtained for central pathology review, the local site pathology report will be reviewed by the TRAVERSE Prostate Adjudication Center at the University of Colorado and the diagnosis and Gleason score reported by the local site pathologist will be used as the endpoint. High grade prostate cancer will be defined as a Gleason score of 4+3 or higher.

Acute urinary retention (AUR) is the inability to voluntarily pass urine, requiring a visit to the emergency department and/ or placement of a catheter to relieve it, ascertained by participant self-report and verified by review of medical record.

An invasive prostate procedure is any surgical procedure on the prostate such as transurethral prostatectomy or open, laser, insertion of prostatic urethral lift (PUL) or incisional prostatectomy for benign prostatic hyperplasia other than a prostate biopsy, ascertained from medical records.

# 12. Baseline Characteristics, Medical History and Study Drug Exposure

## 12.1. Baseline Characteristics

In presenting the FAS we will obtain data from the parent TRAVERSE trial. For presentation of subcohorts (e.g. in subgroup analyses) additional tabular displays may be conducted. Data collected in this sub-study will be documented using summary tables. Statistics for continuous variables will include mean, median, standard deviation, minimum, maximum, and sample size for each treatment group, and two-sided 95% confidence intervals of the mean difference between the treatment groups. Binary variables will be described with frequencies, percentages, and two-sided 95% confidence intervals of the difference in percentages between treatments.

Medical history will mirror the presentation in the parent TRAVERSE trial, and will include at minimum history of depression, CV history; nicotine and alcohol use; and testosterone use.

## 12.2. Report of Treatment Exposure and Compliance

Compliance statistics will be obtained from the parent trial. Where additional displays are necessary (e.g. in exploratory subgroup analyses), we may compute additional summaries relevant to this sub-study. Continuous summaries of subjects' total duration of treatment with study drug may be generated for multiple time intervals (i.e. cumulative to 1- and 2-year follow-up).

# 13. Analysis of Endpoints

All analyses will classify individuals according to their randomized assignment.

**It is acknowledged here that the total number of events observed may not support all analyses (e.g. regression models) described below. In the case that models are not practicable, descriptive alternatives will be employed.**

## 13.1. Assessment for Potential of Ascertainment Bias

A major goal of this substudy will be to capture the incidence rates and rate ratios quantifying the absolute and relative risks of new prostate cancers in the testosterone and placebo arms of the TRAVERSE trial. Owing to the manner by which cancers will be diagnosed - via biopsy, which in turn may be suggested by post-randomization increase in prostate-specific antigen (PSA) - there is the potential for disproportionate capture of indolent cancers in the testosterone arm. Naïve estimation of rates may therefore provide spurious evidence of increased risk in

testosterone arm. Some potential sources of bias, and potential strategies to deal with them, are briefly discussed here

**1. Increase in PSA in testosterone arm, relative to placebo, owing strictly to administration of testosterone.** We anticipate greater and more rapid increase in PSA levels in T arm observable at 3 and 12 months visits than in the placebo arm. We anticipate this increase will be apparent at 3 months and the increase will have leveled off by 12 months, and the rate of PSA increase thereafter will be similar in the two arms, provided cancer risk is similar in the two arms. We hypothesize that continued differentiation in the rate of increase in PSA beyond one year may be evidence of a higher risk of prostate cancer incidence.

**Approach:** At each PSA assessment time point, we will estimate median change from baseline and related descriptive statistics by arm. A linear model will be used to estimate the slope of longitudinal PSA for each treatment group allowing for individual random effects and differing slopes before and after Year 1. We will also graphically display the data over time. We will test whether the slope of PSA varies between the first 12 months and the subsequent intervention period, and if the pattern is different by arm.

**2. The T group will have more biopsies than the Placebo group within the first year**

We expect that exposure to T will cause more men on that arm to cross the PSA threshold (4.0 ng/ml or velocity criterion) and therefore to be recommended for biopsy. The cause for the increase in PSA is likely to be due to one of two causes:

a. An androgen-driven increase in PSA, or
b. Stimulation of pre-existing prostate cancers, i.e., exposure to T via its impact on PSA will identify men who had undetected prostate cancer at study entry.

**Approach**: We will use the PCPT risk calculator to quantify the baseline risk of prostate cancer and high grade disease for all men entering the trial based on PSA, DRE status, age, race, family history, and prior prostate biopsy status. We will compare patterns of baseline risk and recommendation for biopsy across the two arms, and then compare the actual incidence of cancer among individuals with biopsies relative to their predicted probabilities. Although post-randomization PSA is expected to be slightly higher in the T arm, the relative pattern of risk at baseline and subsequent biopsy recommendation will be the same for both arms conditional on PSA level.

**3. There may be more prostate cancers diagnosed in the T arm during the first year compared to Placebo arm, but we anticipate that the incidence will be similar thereafter.**

Owing to the greater rate of biopsy resulting from elevated PSA levels, we expect a few more prostate cancers in that arm during the first 12 months compared to placebo. However, we

expect the rate of prostate cancer (where the number of men who have a biopsy is the denominator) will be comparable between the two arms after year 1.

**Approach:** We will descriptively report on men whose PSA and DRE screening data would suggest that they may be candidates for prostate biopsy, with emphasis on age, race and ethnicity, family history of PCa, PSA at baseline, and the PCPT risk calculator. We will report the number of men who undergo biopsy and the result of that test including grade and stage of disease if positive at 3 months, 1 year and subsequent years by treatment arm. We will plot the cumulative incidence of prostate cancer and high-grade disease by treatment arm over the duration of follow-up.

We will likely see a greater number of biopsies in the testosterone group during the first 12 months, but the biopsy rates will be similar thereafter. If testosterone does not induce new prostate cancers, the curve should look quite a bit like the high-grade (artifactually-caused) cancer curve seen in PCPT.

### 13.2. Primary Analysis

Analyses of high grade prostate cancer occurrence will utilize a proportional hazards regression model for discrete time. The estimated effect of testosterone and its 95% two-sided confidence interval will be extracted from the model adjusted for prior CVD event. Statistical tests comparing two treatment arms will be supported by log-rank test and Kaplan-Meier estimates of the incidence function (cumulative event rates over time) obtained for each intervention group. Potential confounding factors will be considered in sensitivity analyses for secondary outcomes; however, these analyses will be contingent on sufficient number of events occurring throughout the study duration. Models will consider the potential for biases, as noted above, and provide this context in data presentation.

As adjunct to these assessments, we will also consider cumulative incidence of biopsies and cancers at 1 year and during the subsequent intervention period, with estimates obtained from the same regression models described above. We will adopt a parallel approach to consider, among those with biopsy in each arm, the rate of total and high-grade cancers.

### 13.3. Analysis of Secondary Endpoints

Secondary endpoint of prostate cancer occurrence (any grade) will be analyzed by proportional hazards regression model for discrete time. The estimated effect of testosterone vs. placebo and its 95% two-sided confidence interval will be derived from the model. Statistical tests comparing two treatment arms will be supported by log-rank test and Kaplan-Meier estimates of the incidence function (cumulative event rates over time) obtained for each intervention group. Competing risk of death will be considered as appropriate.

Analyses of other time to event outcomes - prostate biopsy, acute urinary retention, start of pharmacologic treatment for lower urinary tract symptoms and invasive prostate surgical procedures for benign prostatic hyperplasia - will be conducted in similar fashion to prostate cancer analyses.

Linear mixed model will be employed to compare effect of testosterone intervention on the change over time in PSA levels, IPSS score, Total and Free Testosterone, DHT, Estradiol and SHBG. Change will be defined as the difference from the baseline value. Models will be adjusted for covariates determined either by inspection to be substantially imbalanced across treatment arms to a degree necessitating adjustment, as determined by combination of numerical imbalance and clinical import of the covariate measure. Models will have as a base set of covariates study visit, treatment effect, visit-by-treatment interaction and baseline value as fixed effects. Random intercept will be included at participants' level. Unstructured covariance matrix will be assumed, however if convergence of the model is not achieved, then a compound symmetry structure will be utilized. Effects of overall change in outcomes over entire treatment period will be calculated as an average score from all visits, and will be extracted, along with two-sided 95% confidence intervals, from the mixed-model framework.

Sensitivity analyses of treatment effect may also be performed where appropriate (i.e. inclusion in the model other stratification factors etc.).

## 14. Changes in this version

Clarified that all analyses in this sub-study will be performed on Safety Analysis Set (not Full Analysis Set). This version also includes data collection conventions and operating procedures.

## 15. References

1. Study protocol of M16-100, Amendment 1. 26 February 2018.
2. Pierorazio PM, Ferrucci L, Kettermann A, et al. Serum testosterone is associated with aggressive prostate cancer in older men: results from the Baltimore Longitudinal Study of Aging. BJU Int. 2010;105(6):824-9.
3. Fowler JE Jr, Whitmore WF Jr. The response of metastatic adenocarcinoma of the prostate to exogenous testosterone. J Urol. 1981;126(3):372-5.
4. Mohr BA, Feldman HA, Kalish LA, et al. Are serum hormones associated with the risk of prostate cancer? Prospective results from the Massachusetts Male Aging Study. Urology. 2001;57(5):930-5.
5. Bhasin S, Singh AB, Mac RP, et al. Managing the risks of prostate disease during testosterone replacement therapy in older men: recommendations for a standardized monitoring plan. J Androl. 2003;24(3):299-311.
6. Roddam AW, Allen NE, Appleby P, et al. Endogenous sex hormones and prostate cancer: a collaborative analysis of 18 prospective studies. J Natl Cancer Inst.2008;100(3):170-83.
7. Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. N Engl J Med. 2004;350(22):2239-46.
8. Schenk JM, Till C, Hsing AW, et al. Serum androgens and prostate cancer risk: results from the placebo arm of the Prostate Cancer Prevention Trial. Cancer Causes Control 2016;27(2):175-182.
9. Morgentaler A, Rhoden EL. Prevalence of prostate cancer among hypogonadal men with prostate-specific antigen levels of 4.0 ng/mL or less. Urology.2006;68(6):1263-7.
10. Calof OM, Singh AB, Lee ML, et al. Adverse events associated with testosterone replacement in middle-aged and older men: a meta-analysis of randomized, placebo-controlled trials. J Gerontol A Biol Sci Med Sci. 2005;60(11):1451-7.
11. Fernández-Balsells MM, Murad MH, Lane M, et al. Clinical review 1: Adverse effects of testosterone therapy in adult men: a systematic review and metaanalysis. J Clin Endocrinol Metab. 2010;95(6): 2560-75.
12. Van Buuren, S., Flexible Imputation of Missing Data. 2012, Chapman & Hall/CRC.

13. Van Buuren, S. and K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 2011. 45(3): p. 67.
14. Ankerst DP, Hoefler J, Bock S, et al. The prostate cancer prevention trial risk calculator 2.0 for the prediction of low-versus high-grade prostate cancer. Urology. 2014;83(6):1362-7.
15. SAS Institute Inc. SAS/STAT® 13.1 User's Guide. Cary, NC: SAS Institute Inc.; 2013.
16. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. American J Epidemiol 2004; 159(7): 702–706.
17. Rivers, C., M.S. Majumder and E.T. Lofgren. Risks of Death and Severe Disease in Patients with Middle East Respiratory Syndrome Coronavirus, 2012–2015. Am J Epidemiol. 2016;184(6): 460-4.

## 16. Schedule of Activities

| Activity | SV1 | SV2 | SV3 | D1 (Baseline) | W 2 | M 1 W 4 | M 3 W 12 | M 6 W 26 | M 9 W 39 (phone) | M 12 W 52 (year 1) | M 15 W 65(phone) | M 18 W 78 | M 21 W 90 (phone) | M 24 W 104 (year 2) | M 27 W 116 (phone) | M 30 W 130 | M 33 W 142 (phone) | M 36 W 156 (year 3) | M 39 W 168 (phone) | M 42 W 182 | M 45 W 194 (phone) | M 48 W 208 (year 4) | M 51 W 220 (phone) | M 54 W 234 | M 57 W 246 (phone) | M 60 W 260/FV (Year 5) | PD | Unscheduled | 30-Day Call |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AE Recording | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Prior/Concomitant Medication | X | X | | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Site Endpoint Questionnaire Form | | | | | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| I-PSS | | X | | X | | | X | | | X | | | | | | | | X | | | | | | | | X | X | | |
| DRE | | X | | | | | | | | X | | | | | | | | X | | | | | | | | X | X | X | |
| PSA | X | | | | | | X | | | X | | | | X | | | | X | | | | X | | | | X | X | | |
| Hematology | | X | | X | | | | X | | X | | X | | X | | | | X | | | | X | | | | X | X | X | |

# 17. Efficacy Analysis Time Windows

## 17.1 For Activity: I-PSS

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M3 | 84 | 2- 224 |
| M12 | 364 | 225 - 728 |
| M36 | 1092 | 729 - 1456 |
| M60 | 1820 | 1457-2180 |
| Final Visit / PD | | 2 to ≤ 2 days after the last dose of study drug |

## 17.2 For Activity: PSA

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M3 | 84 | 2 - 224 |
| M12 | 364 | 225 - 546 |
| M24 | 728 | 547 - 910 |
| M36 | 1092 | 911 - 1274 |
| M48 | 1456 | 1275 - 1638 |
| M60 | 1820 | 1639 - 2180 |
| Final Visit / PD | | 2 to ≤ 2 days after the last dose of study drug |

### 17.3 Time Windows for Time to Event Data

| Scheduled Visit | Nominal Day ( Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1/Baseline[a] | 1 | ≤ 1 |
| Week 2 | 14 | 2 to 21 |
| Week 4 | 28 | 22 to 56 |
| Week 12 | 84 | 57 to 133 |
| Week 26 | 182 | 134 to 227 |
| Week 39 | 273 | 228 to 318 |
| Week 52 | 364 | 319 to 409 |
| Week 65 | 455 | 410 to 500 |
| Week 78 | 546 | 501 to 588 |
| Week 90 | 630 | 589 to 679 |
| Week 104 | 728 | 680 to 770 |
| Week 116 | 812 | 771 to 861 |
| Week 130 | 910 | 862 to 952 |
| Week 142 | 994 | 953 to 1043 |
| Week 156 | 1092 | 1044 to 1134 |
| Week 168 | 1176 | 1135 to 1225 |
| Week 182 | 1274 | 1226 to 1316 |
| Week 194 | 1358 | 1317 to 1407 |
| Week 208 | 1456 | 1408 to 1498 |
| Week 220 | 1540 | 1499 to 1589 |
| Week 234 | 1638 | 1590 to 1680 |
| Week 246 | 1722 | 1681 to 1771 |
| Week 260 (Final Visit) | 1820 | 2 to ≤ 2 days after last dose of study drug |

a. Day of first dose of double-blind study drug.

b. The last value within the window will be used to define Final.

# Statistical Analysis Plan

## Diabetes Sub-Study

## Study M16-100

## The Efficacy of Testosterone Replacement Therapy in Progression from Prediabetes to Diabetes in Hypogonadal Men

## Date: 01 January 2023

## Version 2.2

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| CV | Cardiovascular |
| DPP | Diabetes Prevention Program |
| FAS | Full Analysis Set |
| HbA1c | Hemoglobin A1c |
| HOMA-IR | Homeostatic Model Assessment of Insulin Resistance |
| IRT | Interactive Response Technology |
| LPL | Lipoprotein Lipase |
| MACE | Major adverse CV events |
| MICE | Multivariate imputation by chained equations |
| MMRM | Mixed model repeated measures |
| NHANES | Nation Health and Nutrition Examination Survey |
| ROC | Receiver Operating Characteristic curve |
| SAP | Statistical analysis plan |
| SAS | Statistical Analysis System |
| SD | Standard Deviation |
| SHBG | Sex Hormone–Binding Globulin |
| TRAVERSE | Testosterone Replacement Therapy for Assessment of Long-Term Vascular Events and Efficacy Response in Hypogonadal Men Study |
| VTE | Venous thromboembolism events |

# 1.0 Introduction

This statistical analysis plan supplement (SAP) provides details of the statistical methods for the analyses of data collected as outlined in the protocol (Amendment 2 dated 13 February 2019[1]) for the Diabetes sub-study of The TRAVERSE Trial (M16-100 trial) and describes the analysis conventions to guide the statistical programming work. The scope of this SAP is limited to only the diabetes sub-study.

All analyses will be performed using SAS Version 9.3 or higher (SAS Institute, Inc., Cary, NC 27513) under the UNIX operating system. The SAP will be signed off before the study database is locked.

# 2.0 Study Background

Prediabetes is a metabolic disorder defined by the American Diabetes Association by the presence of fasting plasma glucose concentration of 100 mg/dl (5.6 mmol/L) or higher but less than 126 mg/dL (7 mmol/L), or a 2-hour glucose concentration during a 75g oral glucose tolerance test of 140 mg/dL (7.8 mmol/L) or higher but less than 200 mg/dL (11.1 mmol/L), or a plasma HbA1c 5.7% or higher but less than 6.5% (1-2). Prediabetes has emerged as a major public health problem in the United States and worldwide (2-4). Fourteen percent of US population has diabetes, and nearly 40% has prediabetes (3). The incidence of diabetes and prediabetes has been increasing in the United States and worldwide (3-5). The NHANES survey found an overall 40% prevalence of prediabetes in a multiracial population of US adults (4-5). The prevalence was higher in men than in women and was higher in older adults than in young adults (4-5). In adults aged 45 to 65 years, the prevalence of prediabetes was 45%, and in adults, 65 years or older, the prevalence was 48% (4-5). Because the TRAVERSE Trial will recruit men, 45 to 80 years of age, and two-thirds of men will likely be over 65 years, we assume that the prevalence

will be higher than the 40% prevalence reported in the general US adult population. Furthermore, the eligibility criteria are intended to select men who are at higher metabolic risk for cardiovascular disease than the general US population. For these reasons, we conservatively assume that the prevalence of prediabetes as well as diabetes in the TRAVERSE sample will be higher than in the general US population.

Prediabetes is associated with many adverse health outcomes and poor health-related quality of life (6-13). The risk of progression to type 2 diabetes, cardiovascular disease, end stage renal disease, neuropathy, dementia, some types of cancer, and overall mortality is higher in persons with prediabetes than in the general population (6-13).

The sensitivity and specificity of fasting glucose, 2-hour oral glucose tolerance test and HbA1c for predicting the risk of type 2 diabetes vary, and many experts recommend the use of more than one test to diagnose prediabetes (1-2, 14-15). Fasting glucose A1c are easier to obtain than oral glucose tolerance test, especially in large epidemiologic studies, randomized trials and in clinical practice. The measurement of HbA1c has several advantages over fasting glucose as well as oral glucose tolerance test; HbA1c can be measured in non-fasting samples, its preanalytical stability is superior to glucose concentrations, and the day-to-day variation in HbA1c is less than that in fasting glucose (1). After adjustment for demographic factors, HbA1c-based definition of prediabetes has higher hazard ratios for chronic kidney disease, cardiovascular disease, peripheral arterial disease, and all-cause mortality than fasting glucose concentration (14). However, HbA1c has been reported to have less diagnostic sensitivity than 75 g oral glucose tolerance test, at least in some populations (1).

Here, in the TRAVERSE Trial, we will use both fasting glucose and HbA1c to ascertain prediabetes because it would not be possible to perform 2-hour glucose tolerance test given

the overall participant burden and cost in the context of a large randomized trial whose primary focus is on major adverse cardiovascular events. We recognize that the pathophysiology of hyperglycemia in individuals who are diagnosed by 75 g oral glucose tolerance test may differ in some individuals from those who are diagnosed by impaired fasting glucose or by HbA1c, although there is substantial overlap in pathophysiologic mechanisms in persons diagnosed to have prediabetes by any of the three tests (15-16). Furthermore, the combined use of fasting glucose concentration and HbA1c has been shown to have better predictive value than either test alone for progression to diabetes using receiver operating characteristic curve (ROC) analysis; the area under the curve for the ROC curve of the model that included both fasting glucose and HbA1c was greater than the models that included fasting glucose alone or HbA1c alone (15).

In the Diabetes Prevention Program, the annual incidence of progression to diabetes in a multiracial diverse US population was 11% among adults randomized to placebo (16). The Diabetes Prevention Program recruited adults, 25 years or older, and included both men and women (16). The mean age of the participants was 51 years, and 68% of participants were women. The incidence of progression to diabetes in male participants of the Diabetes Prevention Program was higher than that in the female participants (12.5% vs 10.3%). The TRAVERSE trial will recruit only men and the eligibility age range in the TRAVERSE Trial will be higher - 45 to 80 years - with an average age closer to 60. Therefore, the incidence of progression to diabetes would be expected to be higher in men enrolled in the TRAVERSE Trial than that in the DPP. We assume conservatively that the annual rate of progression from prediabetes to diabetes in the TRAVERSE Trial would be at least 12.5%. Thus the cumulative incidence of diabetes over the average 3.25 years of follow-up in the TRAVERSE Trial would be expected to be at least 35.5%.

Several lines of evidence, described below, support the hypothesis that testosterone replacement therapy might retard progression from prediabetes to diabetes. Many studies have reported that the men with diabetes and metabolic syndrome have lower circulating concentrations of total and free testosterone than men in the general population after adjusting for age (19-28). The men with pre-diabetes also have a higher prevalence of low circulating total testosterone concentrations than men who do not have prediabetes (29).

As men age, their testosterone levels decline and fat mass increases. In epidemiologic studies, serum testosterone levels are consistently associated negatively with fat mass, particularly visceral fat mass (29-32). Testosterone replacement in young and older hypogonadal men is associated with a reduction in whole body, subcutaneous as well as visceral fat mass, and inhibition of uptake of labeled triglycerides and enhanced lipid mobilization in visceral fat (29-32). The experimental induction of androgen deficiency by administration of GnRH agonists or GnRH antagonists in young men is associated with an increase in total fat mass (33). Mechanistically, androgen deficiency is associated with increased lipoprotein lipase activity, resulting in increased fatty acid uptake and triglyceride storage in adipocytes (33). Androgen deficiency also is associated with decreased lipid oxidation rates (33). Furthermore, testosterone is an important regulator of mesenchymal progenitor cell differentiation; it inhibits the differentiation of mesenchymal multipotent progenitor cells into adipocytes (34-35). Testosterone also inhibits the differentiation of preadipocytes into adipocytes through activation of Wnt signaling pathway (34). Through the mediation of these multiple mechanisms, testosterone replacement therapy would be expected to decrease whole body and visceral fat mass.

The relation of testosterone and insulin sensitivity in men is complex and the effects of testosterone replacement therapy on measures of insulin sensitivity have been inconsistent across studies. Epidemiologic studies have reported a negative association between low

testosterone levels, insulin resistance and diabetes risk. One important confounding factor in a number of epidemiological studies has been the influence of sex hormone– binding globulin (SHBG) concentrations on the measured testosterone concentrations. Obese men have lower SHBG concentrations and lower free testosterone concentrations than those who are not obese (28). However, a mendelian randomization study that used data from the United Kingdom Biobank Study found that the effects of genetically determined testosterone levels differ between men and women; higher genetically determined testosterone levels are associated with reduced risk of type 2 diabetes in men, but with increased risk of type 2 diabetes and polycystic ovary syndrome in women (36).

Surgical castration in rats impairs insulin sensitivity; testosterone replacement reverses this derangement (37). However, high doses of testosterone impair insulin sensitivity in castrated rats (37). These data suggest that testosterone effects on insulin sensitivity are biphasic; both low and supraphysiologic testosterone concentrations are associated with suboptimal insulin sensitivity. Marin et al (38) reported that testosterone supplementation of middle-aged men with truncal obesity and low-normal testosterone levels is associated with a reduction in visceral fat volume, serum glucose concentration, and an improvement in insulin sensitivity, suggesting that testosterone is an important regulator of regional fat metabolism. Androgens also increase insulin-independent glucose uptake (39) and modulate LPL activity. The induction of androgen deficiency in men, such as that observed after initiation of androgen deprivation therapy in men with prostate cancer, is associated with induction of insulin resistance and increased risk of developing diabetes mellitus (40). Similarly, abrupt cessation of testosterone replacement therapy in young healthy hypogonadal men is associated with worsening of insulin resistance (41). However, the effects of testosterone replacement therapy on insulin sensitivity in middle-aged and older men with low normal or slightly low testosterone levels have been inconsistent across studies.

Several randomized, placebo-controlled trials of testosterone have been conducted in men with diabetes, although the results of only one such trial have been published. The TIMES2 Study was a randomized trial in which men with type 2 diabetes and/or metabolic syndrome were randomized to either 2% testosterone gel or placebo gel for 6-months (42). Sexual function and plasma lipids improved. Homeostasis Model Assessment of Insulin Resistance (HOMA-IR), a marker of insulin resistance, improved in men who were assigned to testosterone arm and who did not have diabetes vs. placebo. Other small open label uncontrolled studies also have reported reduction in fat mass, and improvements in insulin resistance and glycemic control in hypogonadal men. However, the small size of these trials, inadequate statistical power, heterogeneity of study populations, and uncontrolled nature of some trials limits the strength of the inferences. The TRAVERSE Trial because of its large sample size and long duration provides an outstanding historical opportunity to determine the effects of testosterone on the progression from prediabetes to diabetes.

A number of interventions, including life style modification, metformin, and pioglitazone have been shown to reduce the incidence of progression from prediabetes to diabetes. Life style interventions are highly efficacious in preventing progression to diabetes but they are difficult to implement and overall compliance with life style interventions is suboptimal in real life clinical practice. While we do not advocate the use of testosterone as a therapeutic intervention either to prevent or to treat diabetes, a knowledge of its efficacy in preventing progression from prediabetes to diabetes could be an important consideration in assessing its benefits to risk ratio and could influence individualized therapeutic decision in a substantial fraction of middle-aged and older hypogonadal men who also have prediabetes.

## 2.1 Objective

The aim of the Diabetes Sub-study of the TRAVERSE Trial is to determine the efficacy of testosterone replacement therapy in preventing progression to diabetes in hypogonadal men who have prediabetes.

### 2.1.1 Primary Aims:

1. The primary aim of the diabetes sub-study is to determine the efficacy of testosterone replacement therapy relative to placebo in preventing progression from prediabetes to diabetes in middle-aged and older hypogonadal men.

### 2.1.2 Secondary Aims:

2. To determine whether testosterone replacement therapy compared to placebo, is associated with a greater decrease in fasting glucose levels in randomized participants who have prediabetes at baseline.
3. To determine whether testosterone replacement therapy, compared to placebo, is associated with greater changes in HbA1c levels in the randomized participants who have prediabetes at baseline.
4. To determine efficacy of testosterone replacement therapy, compared to placebo, in inducing glycemic remission among participants with diabetes at baseline.
5. To determine whether testosterone replacement therapy, compared to placebo, is associated with greater changes from baseline in HbA1c levels in randomized participants who have diabetes at baseline.
6. To determine whether testosterone replacement therapy, compared to placebo, is associated with greater changes from baseline in fasting glucose levels in randomized participants who have diabetes at baseline.

7. To determine whether testosterone replacement therapy compared to placebo, is associated with a greater decrease in fasting glucose levels in all randomized participants.
8. To determine whether testosterone replacement therapy, compared to placebo, is associated with greater changes from baseline in HbA1c levels in all randomized participants.
9. To determine whether testosterone replacement therapy, compared to placebo, is associated with greater changes from baseline in Total and Free Testosterone, DHT, Estradiol and SHBG levels in all randomized participants and in people with prediabetes and diabetes at baseline.

### 2.1.3 Hypotheses:

**Primary Hypothesis**

1. Testosterone replacement therapy of middle-aged and older hypogonadal men with prediabetes would be associated with a significantly lower rate of progression to type 2 diabetes mellitus such that a smaller proportion of men randomized to testosterone replacement therapy would develop diabetes than that randomized to placebo.

**Secondary**

1. Testosterone replacement therapy relative to placebo will be associated with significantly greater decrease in the fasting glucose levels in the randomized participants, who have prediabetes at baseline.
2. Testosterone replacement therapy relative to placebo will be associated with significantly greater decrease in the HbA1c levels in the randomized participants, who have prediabetes at baseline.
3. Testosterone replacement therapy relative to placebo will be associated with significantly higher rates of glycemic remission in participants with diabetes at baseline.

4. Testosterone replacement therapy, compared to placebo, will be associated with greater changes from baseline in HbA1c levels in randomized participants who have diabetes at baseline.
5. Testosterone replacement therapy, compared to placebo, will be associated with greater changes from baseline in fasting glucose levels in randomized participants who have diabetes at baseline.
6. Testosterone replacement therapy, compared to placebo, will be associated with a greater decrease in fasting glucose levels in all randomized participants.
7. Testosterone replacement therapy, compared to placebo, will be associated with greater decrease from baseline in HbA1c levels in all randomized participants.

## 2.2 Study Design

### 2.2.1 Study Design and Design Diagram

The diabetes sub-study will be nested within the parent trial. The data for fasting glucose and HbA1c will be analyzed at the end of the trial for the proposed analyses. Additional serum samples will be stored for biomarker analyses at the end of the trial. All the outcome variables, such as MACE, VTE, and thrombotic stroke, are being collected as a part of the parent trial. Thus, the diabetes sub-study will impose no additional burden on the study staff or the participants.

The TRAVERSE parent trial is a Phase 4, randomized, double-blind, placebo-controlled, multicenter study of topical TRT in symptomatic hypogonadal men with increased risk for CV disease. The initial planned study enrollment is approximately 6,000 subjects based on the projected timing when 256 MACE will occur under initial assumptions of the annual event rate, subject accrual rate, and study discontinuation rate. There will be approximately 400 sites in North America and possibly Puerto Rico. An Interactive Response Technology (IRT) system will randomize subjects to receive either topical testosterone gel or placebo

gel in a 1:1 ratio. Randomization will be stratified by pre-existing CV disease (Yes/No). Titration of testosterone dose will occur in subjects receiving active testosterone, while sham dosage titrations will occur in subjects receiving placebo gel via the non-blinded central IRT system. The Screening Period is up to 50 days prior to first dose of study drug. Once subjects meet all of the eligibility criteria during Screening, they will be randomized (1:1 ratio) to active study drug or placebo and will be followed until the study ends. Importantly, randomized subjects who elect to discontinue study drug will also be followed until the study ends unless the subject withdraws from the study completely (withdrawal of informed consent). Subjects who discontinue study drug will still be asked to follow their regularly scheduled protocol visits. Subjects who interrupt study drug will be allowed to restart study drug at any time.

### 2.2.2 Intervention and Intervention Duration

The participants will be randomized to either 1.62% transdermal testosterone gel (Androgel) applied topically daily or to a matching placebo gel.
Maximum follow-up duration is expected to be 5 years and minimum follow-up duration is expected to be 1.5 years. Thus, the median duration of follow-up is expected to be ~3.25 years assuming uniform accrual.

### 2.2.3 Variables Used for Stratification at Randomization

Randomization will descend from the parent trial; no additional randomization or stratification will be imposed in this sub-study. In the parent trial, randomization will be stratified by pre-existing CV disease (Yes/No). Analyses in this diabetes sub-study will acknowledge stratified randomization.

### 2.2.4 Eligibility Criteria

The diabetes sub-study will be nested within the main TRAVERSE trial. All participants in the diabetes sub-study must meet all the inclusion criteria and none of exclusion criteria for the main trial. Additionally, they must also meet the following eligibility criteria to qualify for participation in the diabetes sub-study (analyses supporting Aims 1-3)

**Eligibility criteria for men with pre-diabetes**

**Inclusion Criteria**

- Does not meet the definition of diabetes (see below)
- HgA1c level between 5.7 and 6.4% or one or more screening or baseline fasting glucose levels between 100 and 125 mg/dL.

**Exclusion Criteria**

- Use of diabetes medication
- A diagnosis of diabetes mellitus

**Eligibility criteria for men with diabetes mellitus (for Aim 4)**

A participant is deemed to have diabetes mellitus if he has:

- HgA1c level equal or higher than 6.5% or two fasting glucose levels above 125 mg/dL
- Current diagnosis of diabetes mellitus
- Receiving pharmacologic treatment for diabetes mellitus

Analyses of change in fasting glucose and HbA1c levels will be performed in all participants eligible for the parent trial, in those deemed to have prediabetes at baseline and in those deemed to have diabetes at baseline.

## 2.3 Outcomes

### 2.3.1 Primary outcomes

- The primary endpoint of the diabetes sub-study is the proportion of subjects in each arm who had prediabetes at baseline progressing to diabetes, defined as HbA1c equal to or higher than 6.5%, initiation of diabetes medication, or two consecutive fasting glucose levels >125 mg/dL, assessed at all available time points after baseline.

### 2.3.2 Secondary outcomes

- Change from baseline in glucose, HbA1c, Total and Free Testosterone, DHT, Estradiol and SHBG levels over the intervention period.
- Proportion of participants with diabetes at baseline who have glycemic remission, assessed at all follow-up visits, defined as HbA1c less than 6.5% or two consecutive fasting glucose consecutive assessments less than 126 mg/dL without current use antidiabetic medications.

## 2.4 Statistical Analyses

### 2.4.1 Statistical Analyses for Primary and Secondary Aims

All analyses in this study will utilize intention to treat principles, i.e., all men randomized to an intervention arm will be analyzed in that arm regardless of their compliance. Descriptive characteristics will be summarized by each treatment group and overall. Summary statistics (N, mean, SD, median, quartile range, minimum and maximum) will be provided for continuous variables and the number and percentage of subjects within each category will be presented for categorical data, based on non-missing observations.

**The Operating Procedures of Data Collection and Analysis.** In case the following data operating procedures do not cover any particular aspect of data analysis or variable definitions in this sub-study Statistical Analysis Plan, the operating procedures of the Parent TRAVERSE Trial will be binding. For participants with duplicate ID in the database, data from the duplicated subjects will be excluded from the corresponding analysis sets in this sub-study. An average number of days in a year will be set as 365.25, and an average number of days in 1 month will be set as 365.25/12 for exposure calculation and time to event analyses. The pre-existing cardiovascular disease will be defined, in accordance with parent trial definition, as an occurrence of one or more of the three conditions: coronary artery disease, cerebrovascular disease or peripheral arterial disease. If the same questionnaire data is collected more than once on the same day, the worst score (the highest or the lowest one, depending on the questionnaire) will be used from that day. If there are multiple assessments around the nominal day, the assessment closest to the nominal day will be used.

**Primary analysis** will be performed on participants with prediabetes at baseline. Analyses for the **primary aim** will compare proportions of participants progressing to diabetes during the intervention period using mixed model repeated measures (MMRM) analysis with effects for visit (categorical variable), treatment and visit by treatment interaction terms. Estimation of the risk of progressing to diabetes in testosterone group relative to placebo will be obtained using log-binomial regression (Bernoulli variance function with log link) and robust variance estimation for standard errors and confidence intervals. Treatment effect and its 95% two-sided confidence interval will be extracted from the model adjusted for stratification factors.

Secondary analyses for change from baseline in fasting glucose, HbA1c, Total and Free Testosterone, DHT, Estradiol and SHBG levels will be executed using mixed effect repeated measures (MMRM) model that includes effects for visit, treatment and visit by treatment interaction terms, where the latter quantify the effects of interest at each measurement. No linearity will be assumed for the effect of time unless this is consistent with exploratory analysis (above). Models will also control for stratification factors and will be adjusted to baseline value. . These analyses will be performed separately in 3 populations: all randomized participants, study participants who met the definition of prediabetes at baseline, and study participants who met the definition of diabetes at baseline. Unstructured covariance matrix will be assumed, however if model does not converge compound symmetry will be used. No linearity will be assumed for the effect of time unless this is consistent with exploratory analysis. Secondary analyses will be clearly labeled and performed with no type I error alpha adjustment. Analyses for continuous outcomes (fasting glucose level and HbA1c) will use data from all timepoints.

Analyses of glycemic remission will be performed on participants with diabetes at baseline using mixed model repeated measures (MMRM) analysis in similar fashion to analysis of primary outcome. Estimation of the risk of diabetes remission in testosterone relative to

placebo will be obtained using log-binomial regression (Bernoulli variance function with log link) and robust variance estimation for standard errors and confidence intervals.

### 2.4.2 Sensitivity analysis in African Americans

We recognize that nonglycemic-related factors, especially genetic variants, can influence the results of diabetes screening using HbA1c tests (46-47). As an example, the individuals carrying the HbA1c-lowering alleles of *HBB*-rs334 or *G6PD*-rs1050828 have a lower prevalence of prediabetes and diabetes defined by HbA1c levels compared with noncarriers, whereas there are no differences in the prevalence of diabetes or prediabetes defined by fasting glucose between carriers and noncarriers. These findings prompted the American Diabetes Association to recommend that hemoglobin variants need to be considered when evaluating the HbA1c tests in African Americans (48). Consistent with these recommendations, we will perform sensitivity analyses in the study participants by race and ethnicity; in these sensitivity analyses, prediabetes will be defined only by fasting glucose levels, and diabetes will be defined by fasting glucose, diagnosis of diabetes, or the current use of diabetes medication. The following definitions will be used for sensitivity analyses:

- Pre-diabetes: participants who are not defined as diabetic (based on below definition) and have two consecutive fasting glucose levels equal or above 100 mg/dL.
- Diabetes: two consecutive fasting glucose levels above 125 mg/dL or diabetes mellitus diagnosis or current use of antidiabetic medications.
- Glycemic remission: two consecutive fasting glucose levels below 126 mg/dL without current use of antidiabetic medications.

## 2.5 Interim Analysis

No interim analysis is planned for this sub-study.

## 2.6 Multiplicity Testing Procedures for Type-I Error Control

Type-I error adjustments for multiple comparisons are not planned for efficacy endpoints, subgroup analyses, supportive analyses or sensitivity analyses for this sub-study. All p-values will be considered nominal.

## 2.7 Missing Data

We have no intention to impute data in this sub-study.

If required (e.g. by referees in publication), multiple imputation of endpoints and covariates by the MICE methodology for the report of diabetes sub-study results may be considered where appropriate (i.e. for data reported in the manuscripts). This method is notable for being able to handle clustering of repeated measures at the participant level, a feature of the design of this trial and sub-study. Under such circumstances we would consider whether to impute data for all sub-studies simultaneously.

# 3.0 Analysis Populations and Important Subgroups

## 3.1 Analysis Population

The Analysis Set pre-specified for diabetes sub-study will be a subset of the FAS (Full Analysis Set) comprising all randomized subjects in the main CV study. Eligible subjects from the main study who meet the definition of prediabetes will be included in the analysis. The FAS for the analysis will include all subjects with prediabetes at baseline.

Subjects will be categorized according to treatment assigned at randomization. The FAS will be mainly used for the summary of subjects' disposition and summary of subjects' demographics and baseline characteristics for the sub-study.

Statistical analyses of treatment effect over time will comprise all subjects from the sub-study, who have baseline and at least one post-randomization measurement.

A modified analysis set for participants censored at any time (and subsequently) that they are deemed treatment noncompliant per the parent trial protocol, will be also considered and used for sensitivity analyses of efficacy endpoints.

## 3.2 Subgroup analyses

The following subgroup analysis will be carried out for the main study and may be also considered for diabetes efficacy sub-study:

- by race
- by age (< 65 year, ≥ 65 years)
- by prior CV disease status (yes, no)
- by baseline total testosterone levels (< 250 mg/dl, ≥ 250 mg/dl)

In addition, sub-group analyses supporting aim 1 and 4 will be performed by race using following modified definitions of prediabetes, diabetes status and glycemic remission (Section 2.4.2.).

Additional sub-group analyses for diabetes sub-study, i.e. by age groups (other than 65 years), may be also carried out as appropriate.

# 4.0 Analysis Conventions

## 4.1 Definition of Baseline

Baseline on each outcome measure will be defined as the last available measurement obtained prior to the first dose of study drug (defined as on or before Day 1) (Protocol Appendix C).

## 4.2 Definition of Final Observation

Final observation for each analysis will be the final assessment affiliated with the relevant visit. For example, for twelve-month assessment of glucose levels, the final observation of each relevant measurement affiliated with the 12-month visit will be included in analyses.

## 4.3 Definition of Visit Windows

Definitions of the visit windows (baseline and on-treatment) are presented in Section 10.0 (Tables 10.1) and schedule of sub-study activities is presented in Section 9.0.

### 4.4 Use of Diabetic Medication

Current use of diabetic medication will be assessed at all exams and will be defined as prior diabetic medication use for baseline visit and use of concomitant medication during follow-up visits.

## 5.0 Demographics, Baseline Characteristics, Medical History and Study Drug Exposure

### 5.1 Baseline Characteristics

Data collected in this sub-study will be documented using summary tables. Statistics for continuous variables will include mean, median, standard deviation, minimum, maximum, and sample size for each treatment group, and two-sided 95% confidence intervals of the mean difference between the treatment groups. Binary variables will be described with frequencies, percentages, and two-sided 95% confidence intervals of the difference in percentages between treatments.

Medical history will mirror the presentation in the parent TRAVERSE trial, and will include at minimum history of depression, CV history; nicotine and alcohol use; and testosterone use.

# 6.0 Summary of Changes

## 6.1 Summary of Changes Between the Previous Version and the Current Version

Clarified definition of prediabetes and diabetes.
Specified the analysis of change in A1C and fasting glucose levels in pre-defined study populations

## 6.2 Summary of Changes in Previous Version

Not applicable

# 7.0 References

1. American Diabetes Association. Classification and Diagnosis of Diabetes. Diabetes Care. 2016 Jan;39 Suppl 1:S13-22.
2. Ferrannini E. Definition of intervention points in prediabetes. Lancet Diabetes Endocrinol. 2014 Aug;2(8):667-75.
3. Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and Trends in Diabetes Among Adults in the United States, 1988-2012. JAMA. 2015 Sep 8;314(10):1021-9.
4. Bullard KM, Saydah SH, Imperatore G, Cowie CC, Gregg EW, Geiss LS, Cheng YJ, Rolka DB, Williams DE, Caspersen CJ. Secular changes in U.S. Prediabetes prevalence defined by hemoglobin A1c and fasting plasma glucose: National Health

and Nutrition Examination Surveys, 1999-2010. Diabetes Care. 2013 Aug;36(8):2286-93.

5. National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States, 2011. Atlanta, GA: Centers for Disease Control and Prevention; 2011. Available at: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf. Accessed November 22, 2011

6. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. Lancet. 2012 Jun 16;379(9833):2279-90

7. Ferrannini E, Massari M, Nannipieri M, Natali A, Ridaura RL, Gonzales-Villalpando C. Plasma glucose levels as predictors of diabetes: the Mexico City diabetes study. Diabetologia. 2009 May;52(5):818-24.

8. Ligthart S, van Herpt TT, Leening MJ, Kavousi M, Hofman A, Stricker BH, van Hoek M, Sijbrands EJ, Franco OH, Dehghan A. Lifetime risk of developing impaired glucose metabolism and eventual progression from prediabetes to type 2 diabetes: a prospective cohort study. Lancet Diabetes Endocrinol. 2016 Jan;4(1):44-51

9. Emerging Risk Factors Collaboration., Di Angelantonio E, Gao P, Khan H, Butterworth AS, Wormser D, Kaptoge S, Kondapally Seshasai SR, Thompson A, Sarwar N, Willeit P, Ridker PM, Barr EL, Khaw KT, Psaty BM, Brenner H, Balkau B, Dekker JM, Lawlor DA, Daimon M, Willeit J, Njølstad I, Nissinen A, Brunner EJ, Kuller LH, Price JF, Sundström J, Knuiman MW, Feskens EJ, Verschuren WM, Wald N, Bakker SJ, Whincup PH, Ford I, Goldbourt U, Gómez-de-la-Cámara A, Gallacher J, Simons LA, Rosengren A, Sutherland SE, Björkelund C, Blazer DG, Wassertheil-Smoller S, Onat A, Marín Ibañez A, Casiglia E, Jukema JW, Simpson

LM, Giampaoli S, Nordestgaard BG, Selmer R, Wennberg P, Kauhanen J, Salonen JT, Dankner R, Barrett-Connor E, Kavousi M, Gudnason V, Evans D, Wallace RB, Cushman M, D'Agostino RB Sr, Umans JG, Kiyohara Y, Nakagawa H, Sato S, Gillum RF, Folsom AR, van der Schouw YT, Moons KG, Griffin SJ, Sattar N, Wareham NJ, Selvin E, Thompson SG, Danesh J. Glycated hemoglobin measurement and prediction of cardiovascular disease. JAMA. 2014 Mar 26;311(12):1225-33.

10. Crane PK, Walker R, Hubbard RA, Li G, Nathan DM, Zheng H, Haneuse S, Craft S, Montine TJ, Kahn SE, McCormick W, McCurry SM, Bowen JD, Larson EB. Glucose levels and risk of dementia. N Engl J Med. 2013 Aug 8;369(6):540-8

11. Huang Y, Cai X, Qiu M, Chen P, Tang H, Hu Y, Huang Y. Prediabetes and the risk of cancer: a meta-analysis. Diabetologia. 2014 Nov;57(11):2261-9.

12. Huang Y, Cai X, Chen P, Mai W, Tang H, Huang Y, Hu Y. Associations of prediabetes with all-cause and cardiovascular mortality: a meta-analysis. Ann Med. 2014 Dec;46(8):684-92.

13. Lee M, Saver JL, Hong KS, Song S, Chang KH, Ovbiagele B. Effect of pre-diabetes on future risk of stroke: meta-analysis. BMJ. 2012 Jun 7;344:e3564.

14. Warren B, Pankow JS, Matsushita K, Punjabi NM, Daya NR, Grams M, Woodward M, Selvin E. Comparative prognostic performance of definitions of prediabetes: a prospective cohort analysis of the Atherosclerosis Risk in Communities (ARIC) study. Lancet Diabetes Endocrinol. 2016 Nov 15. pii: S2213-8587(16)30321-7.

15. Sato KK, Hayashi T, Harita N, Yoneda T, Nakamura Y, Endo G, Kambe H. Combined measurement of fasting plasma glucose and A1C is effective for the prediction of type 2 diabetes: the Kansai Healthcare Study. Diabetes Care. 2009 Apr;32(4):644-6.

16. Diabetes Prevention Program Research Group. HbA1c as a predictor of diabetes and as an outcome in the diabetes prevention program: a randomized clinical trial. Diabetes Care. 2015 Jan;38(1):51-8

17. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM; Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002 Feb 7;346(6):393-403.

18. Diabetes Prevention Program Research Group. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. Lancet Diabetes Endocrinol. 2015 Nov;3(11):866-75.

19. Dhindsa, S. et al. Frequent occurrence of hypogonadotropic hypogonadism in type 2 diabetes. *J Clin Endocrinol Metab* **89**, 5462-8 (2004).

20. Haffner SM, Shaten J, Stern MP, Smith GD, Kuller L. Low levels of sex hormone-binding globulin and testosterone predict the development of non-insulin-dependent diabetes mellitus in men. MRFIT Research Group. Multiple Risk Factor Intervention Trial. *Am J Epidemiol*. 1996;143(9):889-97.

21. Stellato RK, Feldman HA, Hamdy O, Horton ES, McKinlay JB. Testosterone, sex hormone-binding globulin, and the development of type 2 diabetes in middle-aged men: prospective results from the Massachusetts male aging study. *Diabetes Care*. 2000;23(4):490-4.

22. Oh J-Y, Barrett-Connor E, Wedick NM, Wingard DL. Endogenous Sex Hormones and the Development of Type 2 Diabetes in Older Men and Women: the Rancho Bernardo Study. *Diabetes Care*. 2002;25(1):55-60.

23. Svartberg J, Jenssen T, Sundsfjord J, Jorde R. The associations of endogenous testosterone and sex hormone-binding globulin with glycosylated hemoglobin levels, in community dwelling men. The Tromso Study. *Diabetes Metab*. 2004;30(1):29-34.

24. Laaksonen DE, Niskanen L, Punnonen K, et al. Testosterone and sex hormone-binding globulin predict the metabolic syndrome and diabetes in middle-aged men. *Diabetes Care*. 2004;27(5):1036-41.

25. Selvin E, Feinleib M, Zhang L, et al. Androgens and Diabetes in Men: Results from the Third National Health and Nutrition Examination Survey (NHANES III). *Diabetes Care*. 2007;30(2):234-238.

26. Ding EL, Song Y, Malik VS, Liu S. Sex Differences of Endogenous Sex Hormones and Risk of Type 2 Diabetes: A Systematic Review and Meta-analysis. *JAMA*. 2006;295(11):1288-1299.

27. Kapoor D, Aldred H, Clark S, Channer KS, Jones TH. Clinical and Biochemical Assessment of Hypogonadism in Men With Type 2 Diabetes: Correlations with bioavailable testosterone and visceral adiposity. *Diabetes Care*. 2007;30(4):911-917

28. Lakshman, K.M., Bhasin, S. & Araujo, A.B. Sex hormone-binding globulin as an independent predictor of incident type 2 diabetes mellitus in men. *J Gerontol A Biol Sci Med Sci* **65**, 503-9 (2010).

29. Ho CH, Yu HJ, Wang CY, Jaw FS, Hsieh JT, Liao WC, Pu YS, Liu SP. Prediabetes is associated with an increased risk of testosterone deficiency, independent of obesity and metabolic syndrome. PLoS One. 2013 Sep 12;8(9):e74173.

30. Woodhouse LJ, Gupta N, Bhasin M, Singh AB, Ross R, Phillips J, Bhasin S. Dose-dependent effects of testosterone on regional adipose tissue distribution in healthy young men. J Clin Endocrinol Metab. 2004 Feb;89(2):718-26.

31. Allan C. A., Strauss B. J., Burger H. G., Forbes E. A. & McLachlan R. I. Testosterone therapy prevents gain in visceral adipose tissue and loss of skeletal muscle in nonobese aging men. J. Clin. Endocrinol. Metab. 93, 139–146 (2008).

32. Allan C. A., Strauss B. J., Burger H. G., Forbes E. A. & McLachlan R. I. Testosterone therapy prevents gain in visceral adipose tissue and loss of skeletal muscle in nonobese aging men. J. Clin. Endocrinol. Metab. 93, 139–146 (2008).

33. Mauras N, Hayes V, Welch S, Rini A,Helgeson K, Dokler M, Veldhuis JD, Urban RJ: Testosterone deficiency in young men: marked alterations in whole body protein kinetics, strength, and adiposity. *J Clin Endocrinol Metab* 83:1886–1892, 1998

34. Gupta V, Bhasin S, Guo W, Singh R, Miki R, Chauhan P, Choong K, Tchkonia T, Lebrasseur NK, Flanagan JN, Hamilton JA, Viereck JC, Narula NS, Kirkland JL, Jasuja R. Effects of dihydrotestosterone on differentiation and proliferation of human mesenchymal stem cells and preadipocytes. Mol Cell Endocrinol. 2008 Dec

35. Singh R, Artaza JN, Taylor WE, Gonzalez-Cadavid NF, Bhasin S. Androgens stimulate myogenic differentiation and inhibit adipogenesis in C3H 10T1/2 pluripotent cells through an androgen receptor-mediated pathway. Endocrinology. 2003 Nov;144(11):5081-8.

36. Ruth KS, Day FR, Tyrrell J, et al. Using human genetics to understand the disease impacts of testosterone in men and women. *Nat. Med.* 2020;26:252-58.

37. Holmang A, Bjorntorp P: The effects of testosterone on insulin sensitivity in male rats. *Acta Physiol Scand* 146:505–510, 1992

38. Mårin P, Ode´n B, Bjo¨rntorp P: Assimilation and mobilization of triglycerides in subcutaneous abdominal and femoral adipose tissue in vivo in men: effects of androgens. *J Clin Endocrinol Metab* 80:239– 243, 1995

39. Hobbs CJ, Jones RE, Plymate SR: Nandrolone, a 19-nortestosterone, enhances insulin-independent glucose uptake in normal men. *J Clin Endocrinol Metab* 81: 1582–1585, 1996

40. Basaria, S., Muller, D.C., Carducci, M.A., Egan, J. & Dobs, A.S. Hyperglycemia and insulin resistance in men with prostate carcinoma who receive androgen-deprivation therapy. *Cancer* **106**, 581-8 (2006).

41. Yialamas, M.A. et al. Acute sex steroid withdrawal reduces insulin sensitivity in healthy men with idiopathic hypogonadotropic hypogonadism. *J Clin Endocrinol Metab* **92**, 4254-9 (2007).

42. Jones, T.H. et al. Testosterone replacement in hypogonadal men with type 2 diabetes and/or metabolic syndrome (the TIMES2 study). *Diabetes Care* **34**, 828-37 (2011). Corona, G. et al. Type 2 diabetes mellitus and testosterone: a meta-analysis study. *Int J Androl* **34**, 528-40 (2011).

43. The NAVIGATOR Study Group. Effect of Nateglinide on the Incidence of Diabetes and Cardiovascular Events. N Engl J Med 2010; 362:1463-1476.

44. Cox DR. Regression models and life-tables. J R Stat Soc [B] 1972;34:187-220.

45. Pocock, S.J., Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. Controlled Clin Trials, 1997; 18: 530-45.

46. Lacy ME, Wellenius GA, Sumner AE, et al. Association of sickle cell trait with hemoglobin A1c in African Americans. JAMA 2017;317:507–515

47. Jee-Young Moon, Tin L. Louie, Deepti Jain, Tamar Sofer, Claudia Schurmann, Jennifer E. Below, Chao-Qiang Lai, M. Larissa Aviles-Santa, Gregory A. Talavera, Caren E. Smith, Lauren E. Petty, Erwin P. Bottinger, Yii-Der Ida Chen, Kent D. Taylor, Martha L. Daviglus, Jianwen Cai, Tao Wang, Katherine L. Tucker, José M. Ordovás, Craig L. Hanis, Ruth J.F. Loos, Neil Schneiderman, Jerome I. Rotter, Robert C. Kaplan, Qibin Qi A Genome-Wide Association Study Identifies Blood Disorder–Related Variants Influencing Hemoglobin $A_{1c}$ With Implications for Glycemic Status in U.S. Hispanics/Latinos. Diabetes Care Sep 2019, 42 (9) 1784-1791; DOI: 10.2337/dc19-0168

48. American Diabetes Association. *Glycemic targets: Standards of Medical Care in Diabetes—2018. Diabetes Care 2018;41(Suppl. 1):S55–S64*

# 8.0 Partial List of Tables with Schedule of Activities

## 8.1 Schedule of Study Assessments

| Assessment | Screening | Baseline | 12-month | 24-month | 36-month | 48-month | 60-month | Final Visit |
|---|---|---|---|---|---|---|---|---|
| Fasting plasma glucose | X | X | X | X | X | X | X | X |
| HbA1c | X | X | X | X | X | X | X | X |

# 9.0 Efficacy Analysis Time Windows

## 9.1 Visit window for diabetes sub-study

| Scheduled Visit | Nominal Day (Study Day) | Time Window (Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M6 | 182 | 92 - 273 |
| M12 | 364 | 274 - 546 |
| M24 | 728 | 547 - 910 |
| M36 | 1092 | 911 - 1274 |
| M48 | 1456 | 1275 - 1638 |
| M60 | 1820 | 1639 - 2180 |
| Final Visit | | 2 to ≤ 2 days after the last dose of study drug |

**Statistical Analysis Plan**

**Study M16-100**

# The Efficacy of Testosterone Replacement Therapy in Correcting Anemia in Middle-aged and Older Hypogonadal Men (The Anemia Substudy)

Date: January 01, 2023

Version 1.8

Table of Contents

List of Abbreviations

| | |
|---|---|
| ACI | Anemia of Chronic Inflammation |
| CBC | Complete Blood Count |
| CKD | Chronic Kidney Disease |
| CRP | C-Reactive Protein |
| CV | Cardiovascular |
| eGFR | Estimated Glomerular Filtration Rate |
| ESC | Executive Steering Committee |
| FAS | Full Analysis Set |
| GFR | Glomerular filtration rate |
| HIS-Q | Hypogonadism Impact of Symptoms Questionnaire |
| IDA | Iron Deficiency Anemia |
| IgA | Immunoglobulin A |
| IgG | Immunoglobulin G |
| IgM | Immunoglobulin M |
| IRT | Interactive Response Technology |
| LDH | Lactate Dehydrogenase |
| MACE | Major Adverse Cardiac Event |
| MCV | Mean Corpuscular Volume |
| MDRD | Modification of Diet in Renal Disease |
| MDS | Myelodysplastic Syndromes |
| SAP | Statistical analysis plan |
| SAS | Statistical Analysis System |

| | |
|---|---|
| SD | Standard Deviation |
| SPEP | Serum Protein Electrophoresis |
| TRAVERSE | Testosterone Replacement Therapy for Assessment of Long-Term Vascular Events and Efficacy Response in Hypogonadal Men Study |
| TRT | Testosterone Replacement Therapy |
| UPEP | Urine Protein Electrophoresis |
| VTE | Venous Thromboembolic Events |

## 1.0 Introduction

This statistical analysis plan supplement (SAP) describes the statistical methods for the analyses of data collected for The Anemia Sub-study of the TRAVERSE Trial (Study M16-100) and provides the analysis plan to guide the statistical programming work. The scope of this SAP is limited to only the Anemia sub-study.

All analyses will be performed using SAS Version 9.3 or later (SAS Institute, Inc., Cary, NC 27513) and/or R version 3.6.0 or later (R Foundation for Statistical Computing, Vienna). The SAP will be signed off before the study database is locked.

## 2.0 Study Background

Although several different definitions of anemia exist in the literature, anemia is currently defined by low hemoglobin levels below 12.7 g/dL using contemporary assays for hemoglobin measurement (1). Anemia is highly prevalent in older adults affecting nearly 10 to 12% of adults, 65 years or older, or 3 to 4 million persons in the United States alone (2-14). The prevalence of anemia rises with advancing age to 20 to 30% among those who are 85 years or older. In about one third of community-dwelling older adults, a clearly definable cause of anemia is not identified; these individuals meet the definition of the unexplained anemia of aging (1).

Unexplained anemia of aging is characterized by mild to moderate decrease in hemoglobin level (hemoglobin levels typically between 10 g/dL to 12.7 g/dL in men) with normocytic red cell indices. The pathophysiology of unexplained anemia of aging is complex and multifactorial, and involves dysregulated erythropoiesis,

reduced response to erythropoietin, reduced stem or progenitor cell proliferative capacity, inhibitory effect of inflammation, and ineffective erythropoiesis. The role of age-related decline in testosterone levels as a contributor to unexplained anemia of aging is incompletely understood.

Unexplained anemia of aging is associated with adverse health outcomes, including impaired quality of life, fatigue, functional limitations, mobility problems, falls, and increased risk of mortality (15-19). Currently, there is no approved therapy for the unexplained anemia of aging.

Testosterone deficiency due to either the disorders of the testis, pituitary and the hypothalamus or to administration of androgen deprivation therapy is associated with a decrease in hemoglobin and hematocrit (20-23). In observational studies, the age-related decline in testosterone levels has been associated with anemia in older men (19-20). Testosterone treatment increases hemoglobin and hematocrit in both young and older men and women (21-23). Erythrocytosis is the most frequent adverse event associated with testosterone treatment (22, 25-26). The older men are more sensitive to the effects of testosterone and exhibit greater increments in hemoglobin and hematocrit than young men (24).

The mechanisms by which testosterone increases hemoglobin and hematocrit are not fully understood. Testosterone stimulates iron-dependent erythropoiesis (27). Testosterone increases iron availability for erythropoiesis through suppression of hepcidin (27-29). Testosterone also stimulates erythropoietin secretion, but the effects of testosterone treatment on serum erythropoietin levels are transient (24, 27, 29). With continued testosterone treatment, serum erythropoietin levels return towards baseline, but are not suppressed below baseline in spite of increased

hemoglobin level (24, 29), suggesting that testosterone treatment alters the set point that regulates the hemoglobin – erythropoietin ratio (29). It is unknown whether testosterone increases the sensitivity of bone marrow erythropoietic progenitors to erythropoietin. Testosterone also appears to have direct effect on the bone marrow hematopoietic progenitor cells (30); it promotes the differentiation of hematopoietic progenitors into common myeloid progenitors. Older adults often suffer from a number of comorbid conditions which are associated with high burden of chronic inflammation, which also may contribute to anemia.

We have shown that testosterone can effectively correct anemia in a mouse model of anemia of inflammation induced by repeated injections of low doses of heat-killed Brucella abortus (31). In this mouse model of anemia of inflammation, testosterone administration reduces ineffective erythropoiesis (31).

Testosterone administration corrects anemia in a preclinical mouse model of aging (32). Although relatively small trials have provided preliminary evidence that testosterone can increase hemoglobin and hematocrit in unexplained anemia of aging (29, 33), a large randomized trial in men with unexplained anemia of aging has not been conducted. The large sample size of the TRAVERSE Trial offers an outstanding opportunity to determine the efficacy of testosterone replacement therapy in correcting unexplained anemia.

In epidemiologic studies, higher levels of hematocrit are associated with increased risk of myocardial infarction, ischemic stroke, and hypertension in both men and women 34-38). However, this relation between hematocrit and cardiovascular risk is complex and varies with age and sex (34). the data on the relation of hematocrit levels with the risk of venous thromboembolic events are less consistent across

studies (39-40). However, none of the previous randomized trials of testosterone treatment has been large enough or long enough to determine whether increases in hematocrit during testosterone treatment are associated with increased risk of myocardial infarction, stroke, or venous thromboembolic events. The TRAVERSE Trial because of its large sample size and rigorous adjudication of MACE, including stroke and venous thromboembolic events offers an outstanding opportunity to address these important questions.

## 2.1 Objective

Some of the proposed aims, described below, will require analyses of the enrolled participants who are anemic at baseline while other aims will require analyses of other groups of participants.

### 2.1.1 Primary Aim:

1. The primary aim of the anemia sub-study is to determine the efficacy of testosterone replacement therapy relative to placebo in correcting anemia in middle-aged and older hypogonadal men.

### 2.1.2 Secondary Aims:

2. To determine whether testosterone replacement therapy in middle-aged and older hypogonadal men with anemia is more efficacious than placebo in increasing the proportion of men whose hemoglobin increases by more than 1.0 g/dL above baseline

3. To determine whether the changes in hemoglobin levels during the intervention are associated with improvements in energy, ascertained using the energy domain of HIS-Q in all randomized participants
4. To determine whether the changes in hemoglobin levels during the intervention are associated with improvements in cognition, ascertained using the cognition domain of HIS-Q in all in all randomized participants
5. To determine whether testosterone treatment, compared to placebo, is associated with greater changes in platelet count, neutrophil, monocyte, lymphocyte counts, red cell counts, white cell counts and RDW.
6. To determine whether increases in hemoglobin, RDW and red cell counts during testosterone treatment are associated with an increased risk of MACE relative to placebo
7. To determine whether increases in hemoglobin and red cell counts during testosterone treatment are associated with an increased risk of having a cerebrovascular event or MACE relative to placebo
8. To determine whether increases in hemoglobin and red cell counts during testosterone treatment are associated with increased risk of venous thromboembolic events (VTE)
9. To determine whether increases in platelet counts during testosterone treatment are associated with increased risk of venous thromboembolic events (VTE), MACE or cerebrovascular events.
10. To determine whether increases in total WBC count, or neutrophil or monocyte counts are associated with increased risk of VTE, MACE or cerebrovascular events
11. To characterize the participant-level factors (such as age, race, baseline body weight and body mass index, smoking, baseline hemoglobin, baseline

creatinine, baseline testosterone level, and increase from baseline in testosterone level) that are associated with the level of increase in hemoglobin, red cell count, and hematocrit

12. To determine whether testosterone administration, relative to placebo, is associated with a differential risk of developing incident anemia among non-anemic men
13. To determine whether testosterone administration, relative to placebo, is associated with a differential change in serum hemoglobin level among non-anemic men

### 2.1.3 Hypotheses:

Primary

1. Among hypogonadal men with anemia, testosterone replacement therapy will be associated with a greater likelihood of correction of anemia relative to placebo.

Secondary

2. Compared to placebo, testosterone replacement therapy will be associated with a greater proportion of middle-aged and older hypogonadal men with anemia (and the entire TRAVERSE trial cohort) whose hemoglobin level increases by more than 1.0 g/dL above baseline.
3. Change from baseline in energy score, measured using the HIS-Q, will be associated with change from baseline in hemoglobin levels.

4. Change in cognition score, ascertained using the cognition domain of HIS-Q will be associated with change in hemoglobin levels.
5. Testosterone treatment relative to placebo will be associated with greater increase in the circulating numbers of platelets, total white blood cells, RDW, red cell count, neutrophils and monocytes but not lymphocytes or other circulating white blood cell types.
6. Change in red cell count, RDW and hemoglobin levels during testosterone treatment will be associated with increased risk of MACE.
7. Change in hemoglobin, red cell count, RDW during testosterone treatment will be associated with increased risk of cerebrovascular accident.
8. Change in red cell count and hemoglobin levels, and RDW during testosterone treatment will be associated with increased risk of venous thromboembolic events (VTE).
9. Change in platelet count during testosterone treatment will be associated with increased risk of VTE, MACE, and cerebrovascular events.
10. Change in total WBC count, and neutrophil and monocyte counts will be associated with increased risk of VTE, MACE, and cerebrovascular events.
11. In analyses of all randomized participants, participant level factors such as age, race, baseline body weight and body mass index, smoking, baseline hemoglobin, baseline creatinine, baseline testosterone level, and increase from baseline in testosterone level will be associated with the level of increase above baseline in hemoglobin, hematocrit, and red cell count.
12. Testosterone administration relative to placebo will be associated with lesser risk of incident anemia in participants who are not anemic at baseline.

13. Testosterone administration relative to placebo will be associated with change in hemoglobin levels among individuals who are not anemic at baseline.
14. Compared to placebo, testosterone replacement therapy will be associated with changes from baseline in Total and Free Testosterone, DHT, Estradiol and SHBG levels in participants not meeting criteria for anemia and in participants meeting criteria for anemia, by study arm and visit

## 2.2 Study Design

The anemia sub-study will be nested within the parent trial. The complete blood counts are being performed in the parent trial for safety monitoring. The data for hemoglobin, hematocrit and red cell indices will be analyzed at the end of the trial for the proposed analyses. Additional serum samples will be stored for biomarker analyses at the end of the trial. All the outcome variables, such as MACE, VTE, and thrombotic stroke, are being collected as a part of the parent trial. Thus, the anemia substudy will impose no additional burden on the study staff or the participants.

### 2.2.1 Study Design and Design Diagram

The TRAVERSE parent trial is a Phase 4, randomized, double-blind, placebo-controlled, multicenter study of topical TRT in symptomatic hypogonadal men with increased risk for CV disease. The initial planned study enrollment is approximately

6,000 subjects based on the projected timing when 256 MACE will occur under initial assumptions of the annual event rate, subject accrual rate, and study discontinuation rate. There will be approximately 400 sites in North America and possibly Puerto Rico. An Interactive Response Technology (IRT) system will randomize subjects to receive either topical testosterone or placebo in a 1:1 ratio. Randomization will be stratified by pre-existing CV disease (Yes/No). Titration of testosterone dose will occur in subjects receiving active testosterone, while sham dosage titrations will occur in subjects receiving placebo gel via the non-blinded central IRT system. The Screening Period is up to 50 days prior to first dose of study drug. Once subjects meet all of the eligibility criteria during Screening, they will be randomized (1:1 ratio) to active study drug or placebo and will be followed until the study ends. Importantly, randomized subjects who elect to discontinue study drug will also be followed until the study ends unless the subject withdraws from the study completely (withdrawal of informed consent). Subjects who discontinue study drug will still be asked to follow their regularly scheduled protocol visits. Subjects who interrupt study drug will be allowed to restart study drug at any time.

The aim of this anemia sub-study is to determine the benefits of testosterone intervention on correcting anemia of aging in middle-aged men, and whether the benefits in older men can be seen with respect to CV occurrence. Furthermore, anemia sub-study will determine whether the changes in hemoglobin levels during the intervention are associated with improvements in energy, ascertained using the energy domain of HIS-Q in all randomized participants.

### 2.2.2 Variables Used for Stratification at Randomization

Randomization will descend from the parent trial; no additional randomization or stratification will be imposed in this sub-study. In the parent trial, randomization will be stratified by pre-existing CV disease (Yes/No). It is expected that in the parent trial at least 30% of the randomized subjects will satisfy inclusion criteria for pre-existing CV disease criteria (secondary prevention), and the remaining 70% will satisfy CV risk factors criteria (primary prevention) combined. Analyses in this PDD sub-study will acknowledge stratified randomization.

### 2.2.3 Eligibility Criteria

All participants enrolled in the parent TRAVERSE trial will be eligible for analyses supporting one or more of the Aims listed above. Participants will be restricted from inclusion in specific analyses on the basis of Aim-specific exclusion criteria given below.

Exclusion Criteria for analyses supporting Aims 1 and 2:

- Hemoglobin at least 12.7 g/dL
- Baseline use of erythropoietic stimulating agents, such as erythropoietin, if known

Exclusion criteria for analyses supporting Aims 12 and 13

- Hemoglobin less than 12.7 g/dL (i.e. participants non-anemic at baseline)

Analyses in Aims 3-11 will be performed on the entire parent trial population.

## 2.3 Outcomes

### 2.3.1 Primary outcome

- Correction of anemia, defined as blood hemoglobin levels at least 12.7 g/dL. This endpoint can vary by study visit.

### 2.3.2 Secondary outcomes

Unless otherwise noted, each outcome can vary by study visit

- Treatment response, as indicated by increase in blood hemoglobin from baseline by more than 1 g/dL at any time.
- Continuous hemoglobin levels following randomization
- Cumulative likelihood of correction of anemia at each timepoint
- Cumulative likelihood of treatment response at each timepoint
- Change in the energy domain score of HIS-Q
- Change in the cognition domain score of HIS-Q
- Change from baseline in platelet count, total white blood cell count, and the numbers of circulating neutrophils, monocytes, and lymphocytes
- Adjudicated MACE, including myocardial infarction and thrombotic stroke, and venous thromboembolic events (VTE)
- Cerebrovascular accident
- VTE
- Cumulative likelihood of incident anemia by time

### 2.3.3 Other measurements

As dictated by availability, pretreatment laboratory results, including red blood cell indices and other components of the complete blood count (CBC) may be employed to characterize participants and their anemia status. Stored serum may be used to perform the following analyses at the end of the trial in men who are deemed anemic, if possible: serum iron, total iron binding capacity, serum ferritin; B12 and folate levels; serum creatinine and estimated GFR using MDRD; LDH and haptoglobin level; CRP, inflammatory cytokines; SPEP/UPEP. Description of these analyses are provided in Appendix 1.

## 2.4 Statistical Analyses and Power

### 2.4.1 Statistical Analyses

**The primary and secondary hypotheses** in the anemia sub-study will be addressed using intention to treat principles. Descriptive characteristics will be summarized by each treatment group. Summary statistics (N, mean, SD, median, quartile range, minimum and maximum) will be provided for continuous variables and the number and percentage of subjects within each category will be presented for categorical data. Exploratory analyses will assess the pattern of change in endpoints and functional form of association between calendar time, events, and continuous measures. Following this, formal analysis will proceed as described below. All statistical estimates will be accompanied by 95% confidence intervals. Hypothesis testing will be conducted at the 0.05 level.

**The Operating Procedures of Data Collection and Analysis.** In case the following data operating procedures do not cover any particular aspect of data

analysis or variable definitions in this sub-study Statistical Analysis Plan, the operating procedures of the Parent TRAVERSE Trial will be binding. For participants with duplicate ID in the database, data from the duplicated subjects will be excluded from the corresponding analysis sets in this sub-study.

An average number of days in a year will be set as 365.25, and an average number of days in 1 month will be set as 365.25/12 for exposure calculation and time to event analyses. The pre-existing cardiovascular disease will be defined, in accordance with parent trial definition, as an occurrence of one or more of the three conditions: coronary artery disease, cerebrovascular disease or peripheral arterial disease.
If the same questionnaire data is collected more than once on the same day, the worst score (the highest or the lowest one, depending on the questionnaire) will be used from that day. If there are multiple assessments around the nominal day, the assessment closest to the nominal day will be used.

**Longitudinal models** will compare risk of incident remission of anemia at all timepoints using a mixed model repeated measures (MMRM) analysis. Estimation of the risk of remission in testosterone relative to placebo will be obtained using log-binomial regression (Bernoulli variance function with log link) and robust variance estimation for standard errors and confidence intervals. If this model fails to converge, the modified Poisson (Poisson likelihood equation with robust variance estimator) will be employed as fit by Generalized Estimating Equations (GEE).

Baseline measurements will be incorporated as outcome measures in a model that includes effects for visit and visit by treatment interaction terms, where the latter

quantify the effects of interest at each measurement. Linearity will not be assumed for the effect of time unless this is consistent with exploratory analysis (above). Models will also control for stratification factors. The treatment effect will be estimated by relative risk and risk differences derived from the interaction terms, along with accompanying confidence intervals, at each measurement of endpoints. Statistical significance will be evaluated using an omnibus test of difference between testosterone and placebo groups at all measurement points simultaneously

Analyses of **secondary aims and outcomes** will proceed in parallel fashion. For continuous endpoints (e.g. HIS-Q subscores and blood biomarker measures), a normal likelihood equation and identify link function will be employed in estimation. For binary endpoints (e.g. MACE, VTE), an approach identical to that described for the primary outcome will be used.

For the purposes of estimating the relative likelihood of correction of anemia and relative likelihood of treatment response at any time point, a discrete-time survival analysis (e.g. proportional hazards model) will be employed. For this analysis, individuals achieving correction of anemia or treatment response at a given visit will no longer be considered at risk of that outcome going forward.

For analyses of incident anemia (conducted among non-anemic participants in support of Aims 12 and 13), methods paralleling those described above for analyses of correction of anemia will be employed.

For other analyses characterizing individuals according to biomarker of anemia, subgroup analysis and regression models contrasting participant subgroups may be employed.

All omnibus tests for treatment effect will be performed using data up to year 3.

### 2.5 Interim Analysis

No interim analysis is planned for this sub-study.

### 2.6 Multiplicity Testing Procedures for Type-I Error Control

Type-I error adjustments for multiple comparisons are not planned for efficacy endpoints, subgroup analyses, supportive analyses or sensitivity analyses for this sub-study.

### 2.7 Missing Data

We have no intention to impute data in this sub-study.

If required (e.g. by referees in publication), multiple imputation of endpoints and covariates by the MICE methodology for the report of anemia sub-study results may be considered where appropriate (i.e. for data reported in the manuscripts). This method is notable for being able to handle clustering of repeated measures at the participant level, a feature of the design of this trial and sub-study. Under such circumstances we would consider whether to impute data for all sub-studies simultaneously.

## 3.0 Analysis Populations and Important Subgroups

### 3.1 Analysis Population

The Analysis Set pre-specified for anemia sub-study will be a subset of the FAS (Full Analysis Set) comprising all randomized subjects in the main CV study.

Eligible subjects from the main study who satisfy the criteria for anemia sub-study will be included in the analysis. The FAS for the analysis of anemia sub-study will include subjects with anemia at baseline.

Subjects will be categorized according to treatment assigned at randomization. The FAS will be mainly used for the summary of subjects' disposition and summary of subjects' demographics and baseline characteristics for the sub-study.

Statistical analyses of treatment effect over time will comprise all subjects from the sub-study, who have baseline and at least one post-randomization measurement.

A modified analysis set for participants censored at any time (and subsequently) that they are deemed treatment noncompliant per the parent trial protocol, will be also considered and used for sensitivity analyses of efficacy endpoints.

Data from eligible measurements will be included, where eligible is defined as being obtained from questionnaires or scores for which at least 80% of items are non-missing.

### 3.2 Subgroup analyses

The following subgroup analysis will be carried out for the main study and may be also considered for anemia efficacy sub-study:

- by race
- by age (< 65 year, ≥ 65 years)
- by prior CV disease status (yes, no)
- by baseline total testosterone levels (< 250 mg/dl, ≥ 250 mg/dl)

To assess potential heterogeneity of effects in hemoglobin changes over time additional sub-group analyses might be considered:

- MCV (mean corpuscular volume) levels: < 100 fL or >= 100 fL,
- MCV < 70 fL
- MCV < 70 fL and red blood cell count 4.35 million per microliter or greater CRP levels: < 10 mg/L or >= 10 mg/L,
- GFR levels: < 30 ml/min or >= 30 ml/min
- RDW levels below or above median

## 4.0 Analysis Conventions

### 4.1 Definition of Baseline

Baseline on each outcome measure will be defined as the last available measurement obtained prior to the first dose of study drug (defined as on or before Day 1) (Protocol Appendix C).

### 4.2 Definition of Final Observation

Final observation for each analysis will be the final assessment affiliated with the relevant visit. For example, for twelve-month assessment of hemoglobin levels, the final observation of each relevant measurement affiliated with the 12-month visit will be included in analyses.

### 4.3 Definition of Visit Windows

Definitions of the visit windows (baseline and on-treatment) are presented in Section 10.0 (Tables 10.1) and schedule of sub-study activities is presented in Section 9.0. These will be harmonized to those of the parent trial, and decisions

made for the parent trial will be considered controlling where they deviate from those described here.

## 5.0 Demographics, Baseline Characteristics, Medical History and Study Drug Exposure

### 5.1 Baseline Characteristics

Data collected in this sub-study will be documented using summary tables. Statistics for continuous variables will include mean, median, standard deviation, minimum, maximum, and sample size for each treatment group, and two-sided 95% confidence intervals of the mean difference between the treatment groups. Binary variables will be described with frequencies, percentages, and two-sided 95% confidence intervals of the difference in percentages between treatments.

Medical history will mirror the presentation in the parent TRAVERSE trial, and will include at minimum history of depression, CV history; nicotine and alcohol use; and testosterone use.

### 5.2 Report of Treatment Exposure and Compliance

Continuous summaries of subjects' total duration of treatment with study drug, among participants recruited in anemia sub-study, will be provided for analyzed time intervals.

Total patient-month of exposure will be calculated by summing the duration of treatment (separately for 1- and 2-year follow-up) for all subjects in the analysis set and dividing this sum by 365.25 (= 1 year). In addition, the number and percentage

of subjects exposed to study drug will be summarized for the following categories of exposure duration: ≤ 1 month, 1 to 3 months, 3 to 6 months, 6 months to 1 year, 1 to 2 years, etc.

Study drug compliance will be computed for anemia sub-study participants, separately for all relevant intervals.

## 7.0 Summary of Changes

### 7.1 Summary of Changes Between the Previous Version and the Current Version

Not applicable.

### 7.2 Summary of Changes in Previous Version

Not applicable

## 8.0 References

1. Beutler E, Waalen J. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? *Blood.* Mar 1 2006;107(5):1747-1750.
2. Izaks GJ, Westendorp RG, Knook DL. The definition of anemia in older persons. *Jama.* May 12 1999;281(18):1714-1717.
3. Merchant AA, Roy CN. Not so benign haematology: anaemia of the elderly. *Br J Haematol.* Jan 2012;156(2):173-185.
4. Salive ME, Cornoni-Huntley J, Guralnik JM, et al. Anemia and hemoglobin levels in older persons: relationship with age, gender, and health status. *J Am Geriatr Soc.* May 1992;40(5):489-496.
5. Chaves PH, Xue QL, Guralnik JM, et al. What constitutes normal hemoglobin concentration in community-dwelling disabled older women? J Am Geriatr Soc 2004;52:1811-6.
6. Guralnik JM, Eisenstaedt RS, Ferrucci L, et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood 2004;104(8):2263---8.
7. Ania BJ, Suman VJ, Fairbanks VF, et al. Incidence of anemia in older people: an epidemiologic study in a well-defined population. J Am Geriatr Soc 1997;45(7):825-31.
8. Artz AS, Fergusson D, Drinka PJ, et al. Mechanisms of unexplained anemia in the nursing home. J Am Geriatr Soc 2004;52(3):423-7.
9. Ble A, Fink JC, Woodman RC, et al. Renal function, erythropoietin, and anemia of older persons: the InCHIANTI study. Arch Intern Med 2005;165(19):2222-7.
10. Joosten E, Pelemans W, Hiele M, et al. Prevalence and causes of anemia in a geriatric hospitalized population. Gerontology 1992;38:111-7.
11. Nilsson-Ehle H, Jagenburg R, Landahl S, et al. Hematological abnormalities and reference intervals in the elderly. A cross-sectional comparative study of three urban Swedish population samples aged 70, 75 and 81 years. Acta Med Scand 1988;224:595-604.

12. Robinson B, Artz AS, Culleton B, et al. Prevalence of anemia in the nursing home: contribution of chronic kidney disease. J Am Geriatr Soc 2007;55:1566-70.
13. Steensma DP, Tefferi A**.** Anemia in the elderly: how should we define it, when does it matter, and what can be done? Mayo Clin Proc. 2007;82:958-66.
14. Thein M, Ershler WB, Artz AS, et al. Diminished quality of life and physical function in community dwelling elderly with anemia. Medicine (Baltimore) 2009;88:107-14.
15. Chaves PH, Ashar B, Guralnik JM, et al. Looking at the relationship between hemoglobin concentration and prevalent mobility difficulty in older women. Should the criteria currently used to define anemia in older people be reevaluated? J Am Geriatr Soc 2002;50:1257-64.
16. Penninx BW, Guralnik JM, Onder G, et al. Anemia and decline in physical performance among older persons. Am J Med 2003;115:104-10.
17. Penninx BW, Pluijm SM, Lips P, et al. et al. Late life anemia is associated with increased risk of recurrent falls. J Am Geriatr Soc 2005;53:2106-11.
18. Penninx BW, Pahor M, Woodman RC, et al. Anemia in old age is associated with increased mortality and hospitalization. J Gerontol A Biol Sci Med Sci 2006;61:474-9.
19. Makipour S, Kanapuru B, Ershler WB**.** Unexplained anemia in the elderly. Semin Hematol. 2008;45:250-4.
20. Ferrucci L, Maggio M, Bandinelli S, et al. Low testosterone levels and the risk of anemia in older men and women. Arch Intern Med 2006;166:1380-8
21. Liverman, CT, Blazer, DG. Testosterone and aging: clinical research directions. Joseph Henry Press, 2004.
22. Bhasin S, Cunningham GR, Hayes FJ, Matsumoto AM, Snyder PJ, Swerdloff RS, Montori VM; Task Force, Endocrine Society. Testosterone therapy in men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 2010;95:2536-59.
23. Snyder PJ, Bhasin S, Cunningham GR, et al. Effects of Testosterone Treatment in Older Men. N Engl J Med. 2016;374:611-624.

24. Coviello AD, Kaplan B, Lakshman KM, Chen T, Singh AB, Bhasin S 2008 Effects of Graded Doses of Testosterone on Erythropoiesis in Healthy Young and Older Men. J Clin Endocrinol Metab 2008;93:914-919.
25. Calof OM, Singh AB, Lee ML, Kenny AM, Urban RJ, Tenover JL, Bhasin S. Adverse events associated with testosterone replacement in middle-aged and older men: a meta-analysis of randomized, placebo-controlled trials. J Gerontol A Biol Sci Med Sci 2005;60:1451-7.
26. Ponce OJ, Spencer-Bonilla G, Alvarez-Villalobos N, Serrano V, Singh-Ospina N, Rodriguez-Gutierrez R, Salcido-Montenegro A, Benkhadra R, Prokop LJ, Bhasin S, Brito JP. The efficacy and adverse events of testosterone replacement therapy in hypogonadal men: A systematic review and meta-analysis of randomized, placebo-controlled trials. J Clin Endocrinol Metab. 2018 Mar 17. doi: 10.1210/jc.2018-00404. [Epub ahead of print]
27. Guo W, Bachman E, Li M, Roy CN, Blusztajn J, Wong S, Chan SY, Serra C, Jasuja R, Travison TG, Muckenthaler MU, Nemeth E, Bhasin S. Testosterone administration inhibits hepcidin transcription and is associated with increased iron incorporation into red blood cells. Aging Cell. 2013 Apr;12(2):280-91.
28. Bachman E, Feng R, Travison T, Li M, Olbina G, Ostland V, Ulloor J, Zhang A, Basaria S, Ganz T, Westerman M, Bhasin S. Testosterone suppresses hepcidin in men: a potential mechanism for testosterone-induced erythrocytosis. J Clin Endocrinol Metab. 2010 Oct;95(10):4743-7.
29. Bachman E, Travison TG, Basaria S, et al. Testosterone induces erythrocytosis via increased erythropoietin and suppressed hepcidin: evidence for a new erythropoietin/hemoglobin set point. Journals of Gerontology. Series A, Biological Sciences and Medical Sciences. 2014;69(6):725-735.
30. Mirand EA, Gordon AS, Wenig J. Mechanism of testosterone action in erythropoiesis. Nature. 1965;206(981):270-2.
31. Guo W, Schmidt PJ, Fleming MD, Bhasin S. Effects of Testosterone on Erythropoiesis in a Female Mouse Model of Anemia of Inflammation. Endocrinology. 2016;157:2937-46.
32. Guo W, Li M, Bhasin S. Testosterone supplementation improves anemia in aging male mice. J Gerontol A Biol Sci Med Sci. 2014;69(5):505-13.
33. Roy CN, Snyder PJ, Stephens-Shields AJ, Artz AS, Bhasin S, Cohen HJ, Farrar JT, Gill TM, Zeldow B, Cella D, Barrett-Connor E, Cauley JA, Crandall JP, Cunningham GR, Ensrud KE, Lewis CE, Matsumoto AM, Molitch ME,

Pahor M, Swerdloff RS, Cifelli D, Hou X, Resnick SM, Walston JD, Anton S, Basaria S, Diem SJ, Wang C, Schrier SL, Ellenberg SS. Association of Testosterone Levels With Anemia in Older Men: A Controlled Clinical Trial. JAMA Intern Med. 2017 Apr 1;177(4):480-490.

34. Gagnon DR, Zhang TJ, Brand FN, Kannel WB. Hematocrit and the risk of cardiovascular disease--the Framingham study: a 34-year follow-up. Am Heart J. 1994 Mar;127(3):674-82.
35. Jin YZ, Zheng DH, Duan ZY, Lin YZ, Zhang XY, Wang JR, Han S, Wang GF, Zhang YJ. Relationship Between Hematocrit Level and Cardiovascular Risk Factors in a Community-Based Population. J Clin Lab Anal. 2015 Jul;29(4):289-93.
36. Kiyohara Y, Ueda K, Hasuo Y, Fujii I, Yanai T, Wada J, Kawano H, Shikata T, Omae T, Fujishima M. Hematocrit as a risk factor of cerebral infarction: long-term prospective population survey in a Japanese rural community. Stroke. 1986 Jul-Aug;17(4):687-92.
37. Brown DW, Giles WH, Croft JB. Hematocrit and the risk of coronary heart disease mortality. Am Heart J. 2001 Oct;142(4):657-63.
38. Panwar B, Judd SE, Warnock DG, McClellan WM, Booth JN 3rd, Muntner P, Gutiérrez OM. Hemoglobin Concentration and Risk of Incident Stroke in Community-Living Adults. Stroke. 2016 Aug;47(8):2017-24
39. Hultcrantz M, Modlitba A, Vasan SK, Sjölander A, Rostgaard K, Landgren O, Hjalgrim H, Ullum H, Erikstrup C, Kristinsson SY, Edgren G. Hemoglobin concentration and risk of arterial and venous thrombosis in 1.5 million Swedish and Danish blood donors. Thromb Res. 2020 Feb;186:86-92.
40. Schreijer AJ, Reitsma PH, Cannegieter SC. High hematocrit as a risk factor for venous thrombosis. Cause or innocent bystander? Haematologica. 2010 Feb;95(2):182-4.

## 9.0 Partial List of Tables with Schedule of Activities

## 9.1 Schedule of Study Assessments

| Assessment | Screening | Baseline | 6 months | 12 months | 24 months | 36 months | 48 months | 60 months |
|---|---|---|---|---|---|---|---|---|
| Complete blood count as in the parent trial | X | X | X | X | X | X | X | X |
| Blood stored for biomarker analyses* | | X | | X | | X | | X |
| His-Q energy and cognition domain** | | X | X | X | X | X | X | X |

* If available

**HIS-Q questionnaire is being administered at these time points as a part of the Anemia Substudy. MACE and thromboembolic events are being recorded as they occur, as part of the parent trial.

## 10.0 Efficacy Analysis Time Windows

## 10.1 Visit window for Anemia sub-study

| Scheduled Visit | Nominal Day<br><br>(Study Day) | Time Window<br><br>(Study Days Range) |
|---|---|---|
| Day 1 | 1 | ≤ 1 |
| M6 | 182 | 92 - 273 |
| M12 | 364 | 274 - 546 |
| M24 | 728 | 547 - 910 |
| M36 | 1092 | 911 - 1274 |
| M48 | 1456 | 1275 - 1638 |
| M60 | 1820 | 1639 – 2180 |
| Final Visit | | 2 to ≤ 2 days after the last dose of study drug |

## Appendix 1. Description of potential biomarker characterizations of anemia

For the purposes of this substudy, participants are classified as anemic if meeting the basic criterion of blood hemoglobin less than 12.7 g/dL. Depending on the availability of additional biomarkers, other analyses may employ characterization of participants by cause of anemia according to the following specification. It is acknowledged that these categories may result in overlapping participant subgroups, which will be acknowledged in analysis and any publication.

Anemic participants may be classified as having a known cause if they had a serum creatinine level of 2.2 mg/dL (to convert to µmol/L multiply by 76.25) or higher (renal insufficiency); either a mean corpuscular volume (MCV) of 105 fL or more and platelet count of 120 000/ mcL (to convert to $\times 10^9$/L multiply by 1) or less or an MCV of 105 fL or more and an absolute neutrophil count less than 1200/ mcL (myelodysplasia); a ferritin level less than 40 ng/mL (iron deficiency) (to convert to pmol/L multiply by 2.247); folate levels less than 3.4 ng/mL (folate deficiency) (to convert to nmol/L multiply by 2.266); vitamin $B_{12}$ less than 200 pg/mL ($B_{12}$ deficiency) (to convert to pmol/L multiply by 0.7378); ferritin levels higher than 500 ng/mL and transferrin saturation less than 50% or ferritin levels higher than 40 ng/mL and a history of a medical condition or medication indicating chronic disease or inflammation (anemia of inflammation); haptoglobin levels less than 14 mg/dL (to convert to mg/L multiply by 10) and MCV greater than 100 fL (hemolytic anemia); and IgG, IgA or IgM levels higher than 1.0 g/dL (plasma cell dyscrasia and/or monoclonal gammopathy).

Iron deficiency anemia: Serum ferritin <50 ng/dL, or % transferrin saturation <20% Recommend < 40 ng/mL

Anemia of chronic inflammation: Serum iron <60 ug/dL, saturation >20%, serum ferritin >50 ng/dl without evidence for iron deficiency (Chart information would be helpful. This is reasonable for lab criteria if we don't have any else. We should summarize the CRP values though in this group to confirm ACI represents patients with ACI). Chronic kidney Disease: eGFR <30 mL/min

Myelodysplastic syndromes: MCV >100 fl, platelet count <120 K/uL, or neutrophil count < 1200 K/uL, not attributable to another cause

Vitamin B12 deficiency: B12 levels <200 pg/mL (if low, further confirmation by measurement of methyl malonic acid

Folate deficiency: Low folate levels (serum folate level < 3.4 ng/L)

Hemolytic anemia: Normocytic or macrocytic anemia associated with elevated LDH and low haptoglobin level

Thalassemia trait: MCV <80 fL and red blood cell count within the normal reference range without iron deficiency

Anemia of aging: Does not meet criteria for IDA, ACI, Mixed IDA/ACI, CKD, and MDS

**Statistical Analysis Plan**

**Study M16-100**

**Fracture Sub-Study**

Date: November 14, 2022

Version 1.3

Version 1.3 – November 14, 2022

## Table of Contents

## List of Abbreviations

| | |
|---|---|
| FAC | Fracture Adjudication Committee |
| TRAVERSE | Testosterone Replacement Therapy for Assessment of Long-Term Vascular Events and Efficacy Response in Hypogonadal Men Study |

## 1.0 Introduction

The Fracture Trial of the Traverse Trial is designed to determine if testosterone treatment of older hypogonadal men will reduce their risk of fractures. All men enrolled in Traverse will be included in the Fracture Trial and will be asked about fractures at every visit; records about reported fractures will be collected. Reported fractures will be adjudicated at the San Francisco Coordinating Center under the auspices of the Fracture Adjudication Committee. This document describes how the resulting data will analyzed statistically at the completion of the trial.

## 2.0 Study Background

As men age, their serum testosterone concentrations decrease[1,2]. Also as men age, their bone density decreases and their incidence of bone fractures increase[3]. Because men who are severely hypogonadal have very low bone mineral density[4], deteriorated trabecular architecture[4] and increased incidence of fractures[5,6], it seems plausible that low testosterone is the cause of the low bone mineral density and increased fractures in older men. In men who were severely hypogonadal due to pituitary or testicular disease, testosterone treatment markedly improved bone mineral density, trabecular architecture and estimated bone strength[7]. In The Testosterone Trials, testosterone treatment of men who were only moderately hypogonadal for no reason other than age also increased bone mineral density and estimated bone strength[8]. The number of participants (788) and duration of observation (one year), however, were too small to draw a conclusion about the effect of testosterone treatment on fracture incidence. The Traverse Trial, which has enrolled more than 5000 hypogonadal men and will be following them for an average of two years each, should allow determination of the effect of testosterone on clinical fracture incidence in this population.

## 3. Objectives

### 3.1 Primary Objective:

- The primary objective of the Fracture Trial is to compare the incidence of clinical fractures in the men treated with testosterone to the incidence in men treated with placebo

**3.2 Secondary Objectives:** Comparison of the two treatment groups regarding

1. Non-high-impact clinical fracture.
2. Clinical fracture in men not taking a medication to treat osteoporosis.
3. Non-high-impact clinical fracture in men not taking a medication to treat osteoporosis.
4. Fracture-free survival.

### 3.3 Exploratory Objectives:

1. Comparison of the two treatment arms with regard to clinical fractures by anatomic site of fracture.
2. Evaluation of the primary outcome adjusting for prognostic factors, such as baseline BMI, medications to treat osteoporosis, and fracture history.
3. Evaluation of the interaction of treatment with any variable showing a significant predictive association with fracture occurrence.
4. Comparison of the two treatment arms with regard to multiple fractures.
5. If testosterone treatment does decrease the incidence of fractures, determine if this effect is associated with the magnitude of the increases in serum testosterone and serum estradiol.

# 4 Study Design

## 4.1 Study Design and Design Diagram

The Traverse Trial is a Phase 4, randomized, double-blind, placebo-controlled, multicenter trial of the effect of testosterone treatment of symptomatic hypogonadal men with increased risk for cardiovascular disease. The trial has enrolled more than 5000 men at more than 300 sites in the United States. Enrollees have been randomized to receive either testosterone or placebo gel in a 1:1 ratio. Randomization is stratified by pre-existing cardiovascular disease (yes or no). The dose of testosterone dose is adjusted in men receiving testosterone gel; sham adjustments are made in men receiving placebo gel. The trial will conclude when 256 major adverse cardiovascular events have occurred.

## 4.2 Variables Used for Stratification at Randomization

Randomization will descend from the parent trial; no additional randomization or stratification will be imposed in this sub-study. In the parent trial, randomization will be stratified by a history of pre-existing cardiovascular disease (yes or no). Analyses in this fracture sub-study will acknowledge stratified randomization.

# 5 Endpoints

## 5.1 Primary Endpoint:

- Time to first clinical fracture

## 5.2 Secondary Endpoints

1. Time to first non-high-impact clinical fracture.
2. Time to first clinical fracture in men not taking a medication to treat osteoporosis.

3. Time to first non-high-impact clinical fracture in men not taking a medication to treat osteoporosis.
4. Fracture-free survival.

### 5.3 Exploratory Endpoints

1. Clinical fractures by anatomic site.
2. Association of prognostic factors, such as baseline BMI, medications to treat osteoporosis, and fracture history, with the primary outcome.
3. Evaluation of the interaction of treatment with any variable showing a significant predictive association with fracture occurrence.
4. Multiple fractures.
5. Relationship of the magnitude of increases in serum testosterone and serum estradiol to the primary outcome.

## 6 Interim Analysis

No interim analysis is planned for this sub-study.

## 7 Multiplicity Analysis

Type 1 error adjustment for multiple comparisons is not planned.

## 8. Analysis Population

Most analyses will utilize the Full Analysis Set (**FAS**) comprising all subjects eligible for analysis in the main TRAVERSE trial. Subjects will be categorized according to treatment assigned at randomization.

## 9. Adjudication of Fractures

Fractures are adjudicated at the San Francisco Coordinating Center under the supervision of the Fracture Adjudication Committee to ensure that fractures are adjudicated blinded, unbiased and consistently. Participants in Traverse are asked at every visit if they have experienced a fracture since the prior visit. If they answer yes, they are asked more questions about the circumstances of the fracture, and the site personnel request documents related to the fracture, such as emergency room records and radiographic reports. These documents are uploaded in the study database. Trained adjudicators evaluate each reported fracture and determine if the reported fracture is confirmed or not confirmed. In some cases, the adjudicator requests additional information, such as the radiographs. Adjudicated fractures are reviewed by the Fracture Adjudication Committee every month.

## 10 Statistical Analyses

### 10.1 Baseline Characteristics

Baseline characteristics will be those from the main Traverse Trial. Baseline characteristics of special interest will be body mass index, smoking and alcohol use and history of prior fractures. Statistics for continuous variables will include mean, median, standard deviation, minimum, maximum, and sample size for each treatment group. Binary variables will be described with frequencies and percentages.

### 10.2 Primary Analysis

The primary outcome is clinical fracture (as confirmed by the Fracture Adjudication Committee (FAC), occurring during the treatment period and defined as a clinical spine or non-spine fracture documented by a positive imaging test and/or surgery, irrespective

of the degree of trauma). Fractures of the sternum, fingers, toes, facial bones and skull will be excluded. Hereinafter, the term "fracture" will be used to mean a confirmed clinical fracture with these exclusions.

The primary analysis will be a time-to-event analysis, considering first confirmed clinical fracture (with exclusions as noted above) as an event, as this will be more powerful than a simple comparison of the proportion of men experiencing a fracture over the course of the study. We shall compare time to first confirmed clinical fracture using a cause-specific Cox proportional hazards model. The model will include a term indicating prior cardiovascular disease status (yes or no), as well as the treatment indicator (testosterone vs placebo). Men who become lost to follow-up or have not experienced a fracture by the end of the study will be censored at time of last contact. The primary analysis will include men in their randomized group without consideration of adherence to treatment (Intent to Treat). The hazard ratio estimating the relative effects in the two treatment groups will be estimated by the cause-specific Cox proportional hazards model, including history of cardiovascular disease as a covariate.

### 10.3 Secondary Analysis

1. Time to first clinical fracture in the subgroup of men who were not taking a medication to treat osteoporosis prior to the fracture. The medications are alendronate, risedronate, ibandronate, zoledronate, bisphosphonates, denosumab, teriparatide, abaloparatide, and romosozumab.
2. Time to first non-high-impact clinical fracture (all clinical fractures excluding the category "severe non-fall trauma"), as adjudicated by the Fracture Adjudication Committee.
3. Time to first non-high-impact clinical fracture in the subgroup of men not taking medications to treat osteoporosis,

4. “Fracture-free survival,” in which death as well as clinical fracture (as defined above) will be counted as an event, will also be a secondary analysis.

These secondary analyses will be performed using cause-specific Cox proportional hazards models, as for the primary outcome, with a history of cardiovascular disease as a covariate.

### 10.4 Exploratory Analysis

We shall perform the following exploratory analyses to expand our understanding of the potential effect of testosterone treatment on fracture risk. All analyses will be adjusted for the stratification factor of history of cardiovascular disease.

1. Time to first clinical fracture not excluding fingers, toes, skull, facial bones and sternum.
2. Time to first clinical fracture not excluding those fractures classified as “uncertain”.
3. Descriptive summaries of fracture rates by anatomic site of fracture will be provided. While we do not anticipate adequate statistical power for tests of site-specific differences, we shall separately compare times to hip fracture, clinical vertebral fracture and major osteoporotic fractures (hip, humerus, wrist and clinical spine) using cause-specific Cox proportional hazards models, including a history of cardiovascular disease as a covariate, as these are the most concerning types of fractures in older populations.
4. Evaluation of the primary outcome using cause-specific Cox proportional hazards analysis that adjusts for prognostic factors in addition to the stratification factor and can include continuous as well as categorical covariates. These will include baseline BMI; a binary variable indicating use, prior to fracture, of medications to treat osteoporosis (alendronate, risedronate, ibandronate, zoledronate, denosumab,

teriparatide, abaloparatide, romosozumab); any fracture after age 45 but prior to entering the trial; and age and baseline serum testosterone concentration as continuous variables. In addition, the same Cox proportional hazards model will be assessed by replacing any fracture after age 45 but prior to entering the trial with separate indicators for nonvertebral fracture and vertebral fracture (both are after age 45 but prior to entering the trial).

5. Tests of interaction of treatment with any variable showing a significant ($P<0.1$) predictive association with fracture occurrence by adding interaction terms into the Cox regression model. These may include, for example, age and current use of bone-strengthening medications. When the interaction is statistically significant, we will perform separate stratified comparisons of time to fracture within subgroups defined by such variables.
6. Comparisons of the number of men in each arm who experience multiple fractures, using a time-to-second fracture analysis adjusting for cardiovascular history.
7. If testosterone treatment does decrease the incidence of fractures, as shown by a significant reduction in fracture incidence in the primary analysis, we shall investigate whether this effect of testosterone is associated with the magnitude of the increases in serum testosterone, free testosterone, estradiol and dihydrotestosterone using the hormone levels measured at progressive intervals during the trial as time-dependent covariates in the cause-specific Cox model for the primary analysis described above.

### 10.5 Sensitivity Analyses

1. Assessment of the potential impact of missing data on fracture incidence resulting from study dropouts and losses to follow-up, using the method of multiple imputation

to estimate the missing outcomes using observed baseline characteristics and outcome measures. These methods assume that the missing data are "missing at random" once observed data are accounted for. We will do further sensitivity analyses under the assumption that the data are "non-ignorable," which will require modeling the missingness mechanism, using shared parameter models and pattern-mixture models. The outcome model will be equivalent to the model for the primary analysis. We will use a time-to-event approach to evaluate the probability of dropout over the course of the study. We will compare coefficient estimates from the shared parameter and pattern-mixture models to estimate from the analysis under the "missing at random" assumption.

2. We shall also assess the potential impact of discontinuation of treatment in several ways. We shall perform an analysis considering only fractures observed no more than three months following discontinuation of treatment. We shall also use the method of inverse probability weighting to adjust for pre- and post-randomization covariates that may be associated both with adherence and fracture risk.

## 12.0 References

1. Harman SM, Metter EJ, Tobin JD, Pearson J, Blackman MR, Baltimore Longitudinal Study of A. Longitudinal effects of aging on serum total and free testosterone levels in healthy men. Baltimore Longitudinal Study of Aging. The Journal of clinical endocrinology and metabolism 2001;86:724-31.

2. Wu FC, Tajar A, Pye SR, et al. Hypothalamic-pituitary-testicular axis disruptions in older men are differentially linked to age and modifiable risk factors: the European Male Aging Study. The Journal of clinical endocrinology and metabolism 2008;93:2737-45.

3. Riggs BL, Wahner HW, Seeman E, et al. Changes in bone mineral density of the proximal femur and spine with aging. Differences between the postmenopausal and senile osteoporosis syndromes. The Journal of clinical investigation 1982;70:716-23.
4. Benito M, Gomberg B, Wehrli FW, et al. Deterioration of trabecular architecture in hypogonadal men. The Journal of clinical endocrinology and metabolism 2003;88:1497-502.
5. Shahinian VB, Kuo YF, Freeman JL, Goodwin JS. Risk of fracture after androgen deprivation for prostate cancer. The New England journal of medicine 2005;352:154-64.
6. Smith MR, Boyce SP, Moyneur E, Duh MS, Raut MK, Brandman J. Risk of clinical fractures after gonadotropin-releasing hormone agonist therapy for prostate cancer. J Urol 2006;175:136-9; discussion 9.
7. Benito M, Vasilic B, Wehrli FW, et al. Effect of testosterone replacement on trabecular architecture in hypogonadal men. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research 2005;20:1785-91.
8. Snyder PJ, Kopperdahl DL, Stephens-Shields AJ, et al. Effect of Testosterone Treatment on Volumetric Bone Density and Strength in Older Men With Low Testosterone: A Controlled Clinical Trial. JAMA internal medicine 2017;177:471-9.

## 1.0 Title Page

# Statistical Analysis Plan

# Study M16-100

# Testosterone Replacement Therapy for Assessment of Long-Term Vascular Events and Efficacy ResponSE in Hypogonadal Men (TRAVERSE) Study

# Date: 11 Apr 2018

# Version 1.0

## 2.0 Table of Contents

## List of Tables

# List of Abbreviations

| | |
|---|---|
| 5-ARI | 5-Alpha Reductase Inhibitor |
| AE | Adverse event |
| AESI | Adverse events of special interest |
| ALT | Alanine aminotransferase |
| AST | Aspartate aminotransferase |
| BMI | Body mass index |
| BSSR | Blinded Sample Size Re-estimation |
| BUN | Blood urea nitrogen |
| CABG | Coronary artery bypass graft |
| CEC | Clinical Events Committee |
| CI | Confidence interval |
| CIF | Cumulative incidence function |
| CRF | Case report form(s) |
| CV | Cardiovascular |
| DBP | Diastolic blood pressure |
| DHT | Dihydrotestosterone |
| DMC | Data Monitoring Committee |
| DVT | Deep vein thrombosis |
| ECG | Electrocardiogram |
| ESC | Executive Steering Committee |
| FAS | Full Analysis Set |
| FV | Final visit |
| GGT | Gamma-glutamyl transferase |
| Hct | Hematocrit |
| Hgb | Hemoglobin |
| HbA1c | Hemoglobin A1c |
| HR | Hazard ratio |
| I-PSS | International Prostate Symptom Score |
| IRT | Interactive Response Technology |
| KM | Kaplan-Meier |
| LLN | Lower limit of normal |
| LTFU | Lost to follow-up |
| MACE | Major Adverse Cardiac Event |

| | |
|---|---|
| MCV | Mean Corpuscular Volume |
| MCH | Mean Corpuscular Hemoglobin |
| MCHC | Mean Corpuscular Hemoglobin Concentration |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MI | Myocardial infarction |
| PCI | Percutaneous coronary intervention |
| PD | Premature discontinuation |
| PE | Pulmonary embolism |
| PSA | Prostate specific antigen |
| PT | Preferred term |
| QTcF | QT interval corrected for heart rate (Fridericia's correction formula) |
| RBC | Red blood cell |
| RD | Risk difference |
| RMST | Restricted mean survival time |
| SAE | Serious adverse event |
| SAP | Statistical analysis plan |
| SAP-S | SAP- supplement |
| SBP | Systolic blood pressure |
| SOC | System organ class |
| TE | Treatment emergent |
| TEAE | Treatment emergent adverse event |
| TRT | Testosterone Replacement Therapy |
| TTE | Time to Event |
| ULN | Upper limit of normal |
| WBC | White blood cell |

# 3.0 Introduction

This statistical analysis plan (SAP) provides details to elaborate statistical methods for data collected as outlined in the protocol (Amendment 1 dated 26 February 2018[1]) for Study M16-100 and describes analysis conventions to guide the statistical programming work. The scope of this SAP is limited to only the main cardiovascular (CV) safety non-inferiority study. The efficacy sub-studies do not fall within the scope of this SAP. Analysis of prostate safety endpoints (including prostate cancer and other prostate events) are not intended to be the part of the main CV safety CSR and hence are also excluded from this SAP. These analyses will be detailed in separate SAP(s).

All analyses will be performed using SAS Version 9.2 or higher (SAS Institute, Inc., Cary, NC 27513) under the UNIX operating system. The SAP will be signed off before the study database is locked.

This SAP will *not* be updated in case of future (administrative or minor) amendments to the protocol unless the changes have any impact on the analysis of study data described here.

# 4.0 Study Background

## 4.1 Objective

The following primary and secondary objectives will be evaluated in the TRAVERSE Study population, which is comprised of middle-aged and older hypogonadal men at risk for CV disease:

**Primary Objective:** To compare the effect of testosterone replacement therapy (TRT) and placebo on the incidence of major cardiac adverse events (MACE). The composite MACE endpoint consists of CV death, non-fatal MI and non-fatal stroke, as adjudicated by the Clinical Events Committee (CEC) of the study.

**Secondary Objectives:**

- To compare the effect of TRT and placebo on an expanded list of CV outcomes including MACE plus coronary revascularization procedures/cardiac percutaneous coronary intervention (PCI) and coronary artery bypass graft (CABG) surgery.
- To determine the effect of TRT on the incidence of high-grade prostate cancer. (Note: Evaluation of the effect on the incidence of prostate cancer will be specified in a separate SAP. Evaluation of this secondary objective is not intended to be part of the main CV safety CSR, and will be reported separately.)

## 4.2 Study Design

### 4.2.1 Study Design and Design Diagram

This is a Phase 4, randomized, double-blind, placebo-controlled, multicenter non-inferiority study of topical TRT in symptomatic hypogonadal men with increased risk for CV disease. The planned study enrollment is to randomize approximately 6,000 subjects to target 256 MACE. The expected duration of the study is approximately 5 years.

The screening period for a subject is up to 50 days prior to randomization. Subjects who fail to meet the selection criteria may be re-screened one time at the discretion of the Investigator once their clinical status changes provided they did not screen fail due to testosterone levels outside the entry criteria. Once subjects meet all of the eligibility criteria during Screening, they will be randomized (1:1 ratio) to active study drug or placebo and will be followed until the study ends. Randomization will be stratified by pre-existing CV disease (Yes/No). All eligible subjects will be started at 40.5 mg of study drug, either active drug or placebo on Study Day 1 (Baseline). Titration of testosterone dose (see Protocol Section 5.5.4) will occur in subjects receiving active testosterone in pre-specified study visits, while sham dosage titrations will occur in subjects receiving placebo gel via the non-blinded central interactive response technology (IRT) system.

Randomized subjects who elect to discontinue study drug will also be followed until the study ends unless the subject withdraws from the study completely (see Protocol Section 5.4.1). Subjects who discontinue study drug will still be asked to follow their regularly scheduled protocol visits specified in protocol Appendix C Study Activities. Subjects who interrupt study drug will be allowed to restart study drug at any time as long as the study is ongoing.

Randomized subjects will return for the Final Visit (FV) (Protocol Appendix C) once the Sponsor closes the study. For randomized subjects who discontinue the study prematurely, Premature Discontinuation (PD) Visit case report forms (CRFs) should be completed at the time the subject withdraws consent (e.g., in-person preferably or via a phone visit) from the study or has a fatal outcome (if this information is available). For lost to follow-up subjects, every attempt should be made to contact the subjects to obtain study related information. For these subjects, the FV CRFs should not be completed until the end of the study (see Protocol Section 5.4.1) so that further study information can be obtained. Subjects meeting the stopping criteria for study drug (see Protocol Section 5.4) will have study drug discontinued and will be followed for all safety events outlined in Protocol Appendix D for the duration of the study (see Protocol Section 6.1.4).

Periodic blinded overall MACE assessments, coupled with the sample size and duration of treatment to date, will be conducted to estimate whether or not the sample size and/or the study duration will need to be modified in order to observe at least 256 MACE. Study data will be monitored by an independent Data Monitoring Committee (DMC) in accordance with standard conduct for clinical outcome studies. Study oversight will occur through an Executive Steering Committee (ESC), which will advise the Sponsor and the independent DMC. Separate Clinical Events Committees (CEC) will be utilized to adjudicate pre-specified CV, prostate and fracture event data during the study. The governance of each of these committees will be covered by separate charters.

### 4.2.2 Variables Used for Stratification at Randomization

Randomization will be stratified by pre-existing CV disease (Yes/No). It is expected that 30% of the randomized subjects will satisfy inclusion criteria for pre-existing CV disease criteria (secondary prevention), and the remaining 70% will satisfy CV risk factors criteria (primary prevention) combined; these proportions will be monitored by the study team. The ESC and Sponsor may decide to cap the cohort of subjects with CV risk factors (i.e., no pre-existing CV disease) if that cohort is found to consistently exceed 70% of the total population enrolled or if the pooled primary event rate falls below projections.

## 4.3 Endpoint

### 4.3.1 Primary Endpoint (Safety)

The primary endpoint of the study, time to MACE, is defined as time from randomization to the first occurrence of any component event of the composite MACE endpoint consisting of CV death, non-fatal myocardial infarction (MI) and non-fatal stroke, as adjudicated by Clinical Events Committee (CEC). Any undetermined death will also be considered as CV death. Time to MACE for subjects who do not experience an event on study will be right-censored at the time of their last available follow-up observation.

### 4.3.2 Secondary Endpoints (Safety)

Secondary CV Safety Endpoint

The secondary CV safety endpoint is defined as time from randomization to first occurrence on any component event of the following (as adjudicated by CEC):

- Non-fatal MI
- Nonfatal stroke
- Death due to CV causes (including undetermined deaths)
- Coronary revascularization procedures/cardiac PCI and CABG

Secondary Prostate Safety Endpoint

- Incidence rate of high grade prostate cancer (Gleason score of 4 + 3 or higher) (Note: Details of the analysis of incidence of high prostate cancer will be specified in a separate SAP. Summaries of this secondary endpoint are not intended to be part of the main CV safety CSR, and will be reported separately.)

### 4.3.3 Tertiary Endpoints (Safety)

The tertiary safety endpoints for the study are:

- Incidence rate of all-cause mortality
- Incidence rate of heart failure events (hospitalization or urgent visit)
- Incidence rate of thromboembolic Events to include deep vein thrombosis (DVT)/ pulmonary embolism (PE)/venous thromboembolism (excluding superficial thrombophlebitis)
- Incidence rate of peripheral arterial revascularization

All CEC adjudicated events will be included in the analysis.

### 4.3.4 Other Safety Endpoints

In addition to the safety endpoints mentioned in Section 4.3.1, Section 4.3.2, and Section 4.3.3, the following safety endpoints will also be analyzed:

- Incidence rate of Adverse events
- Incidence rate of lab abnormalities
- Change from baseline in vital signs parameters including systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse rate, temperature, body mass index (BMI) and weight
- Change from baseline in Electrocardiogram (ECG) parameters
- Change from baseline in Serum testosterone level, Serum free testosterone level, Serum Dihydrotestosterone (DHT) level and Serum Estradiol level

- Change from baseline in International Prostate Symptom Score (I-PSS) score

### 4.3.5 Efficacy Endpoint

No efficacy endpoints from sub-studies are covered in this main SAP. Separate SAP(s) will accompany this SAP at a later date describing the full analyses for the efficacy sub-studies.

### 4.3.6 Pharmacological Endpoint

There are no pharmacological endpoints planned in this study.

## 4.4 Sample Size Justification

This non-inferiority study plans to observe a total of 256 primary composite events (i.e., MACE) to rule out a hazard ratio (HR) of 1.5 at the 95% (2-sided) upper confidence limit (i.e., 1-sided alpha = 2.5%) with 90% power on the estimated annual placebo event rate of 1.5%, an accrual period of 3.5 years, and an annualized lost to follow-up (LTFU) rate of 2%. A total of approximately 5,400 subjects (2,700 per treatment arm) are needed to observe the 256 required events for the primary analysis.

However in order to achieve similar power for the principal sensitivity analysis (i.e., analysis based on censoring subjects without events after 365 days post last dose), approximately 6,000 subjects (in 3.5 year accrual period) will be needed assuming a treatment discontinuation rate is 20% in the first year, and 10% in the second year and thereafter.

Therefore, the study is planned to enroll approximately 6,000 subjects (3,000 per treatment arm), and the study will be stopped and analysis will be conducted after 256 MACE are observed for the principal sensitivity analysis. The study duration is projected to be 5.2 years under the alternative hypothesis (True HR = 1.0, i.e., no increased risk with TRT) and 4.4 years under the null hypothesis (True HR = 1.5).

### 4.4.1 Blinded Sample Size Re-Estimation (BSSR)

Certain design parameters may be observed to be different during the course of the study from their assumed values during the study design process, most notably the assumed accrual rate, the annualized placebo event rate, and LTFU rate. Therefore, a periodic **blinded** review of the pooled study data will occur during the course of the study (e.g., after 4,500 subjects have been randomized or at approximately 2.5 years from the first subject enrolled) to reassess these parameters and consider their impact on other planned study duration and sample size.

The accrual rate and pooled event LTFU and treatment discontinuation rates will be evaluated against the original estimates used in the study design. The total number of subjects that will be enrolled for the study may be adjusted according to the observed pooled event rate, LTFU rate, as well as the accrual rate (the accrual period may need to be adjusted too) in order to obtain the target 256 events within the planned study duration (i.e., approximately 5 years assuming no increased risk with TRT). The aforementioned adjustments and final sample size determinations will take this consideration into account. Detailed methodology of the BSSR is outlined below.

In general, for time-to-event analysis, the sample size needed to achieve a certain number of events is a function of event rate (p), LTFU rate (d), accrual period (r), total study duration (t), and total required number of events (e). Thus, we can write the below formula for the sample size calculation, assuming constant hazard.

$$n = f(p, d, r, t, e)$$

In the study design, we assumed

$p = 1.5\%$ per year placebo arm, $1.5\%$ for TRT arm when $\text{HR} = 1.0$

$d = 2\%$ per year

$r = 4$ years

$t = 5.2$ years when HR $= 1.0$

$e = 256$ events

Assume the BSSR is conducted at time $t_1$ from the first subject enrolled, and the following notations:

$n_1$: number of subjects enrolled at $t_1$

$e_1$: number of events observed at $t_1$

$p_1$: pooled event rate observed at $t_1$

$d_1$: pooled LTFU rate at $t_1$

For those subjects among $n_1$ who haven't had events yet, we can predict the additional number of events ($e_1'$) that will occur in the remaining of the study (i.e., in the interval from $t_1$ to $t$) based on observed event rate ($p_1$) and LTFU rate ($d_1$).

The additional sample size needed for the remaining of the study can be estimated as

$$n_2 = f(p_1, d_1, r_2, t_2, e_2)$$

Where $e_2 = 256 - e_1 - e_1'$; $t_2 = t - t_1$; and $r_2$ (accrual duration in the remaining of the study) can be adjusted if necessary.

Therefore, the total sample size will be $n = n_1 + n_2$.

## 4.5 Interim Analysis

No interim analysis is planned for the study.

### 4.6 Multiplicity Testing Procedures for Type-I Error Control

Type I error adjustments for multiple comparisons are not planned for safety endpoints, subgroup analyses, supportive analyses or sensitivity analyses for this study.

### 4.7 Missing Data Imputation

No imputation of missing data is planned for the primary and secondary time-to-events of this study. If the event of interest is not observed for a subject in the study, such as when the subject is lost to follow-up, the endpoint will be right censored at the last adequate observation time when occurrence of an event can be ruled out.

## 5.0 Analysis Populations and Important Subgroups

### 5.1 Analysis Population

The following data sets will be used for the analyses of all safety endpoints for this study:

The Full Analysis Set (FAS) will comprise all randomized subjects and will be used mainly for the summary of subjects' disposition and subjects' demographics and baseline characteristics for the study. Subjects will be categorized to the treatment arms according to the treatment assigned at randomization.

The Safety Set will comprise all randomized subjects who receive at least one dose of study drug (TRT or placebo). Unless otherwise specified, the Safety Set will be used for the analysis of all safety endpoints of the study, in particular, the primary and secondary endpoints for this study. Subjects will be categorized to the treatment arms according to first treatment received.

### 5.2 Subgroup

Following subgroup analysis will be carried out for the main study:

- by race
- by age ($< 65$ year, $\geq 65$ years)

- by prior CV disease status (yes, no)

Additional sub-group may be carried out and that will be specified in SAP-S.

# 6.0 Safety Analyses

## 6.1 General Considerations

The analyses of all study endpoints will be primarily based on the Safety Set. Analyses of these endpoints based on the FAS will only be supportive in nature. Summaries of the incidence rates of adverse events (AEs), laboratory parameters, vital signs parameters and electrocardiogram (ECG) parameters will be provided only using the safety set. Unless otherwise specified, all CEC adjudicated cardiac events will be included in the analysis. For all other safety endpoints (e.g., AEs, lab values, vital sign values and ECG values), only the treatment emergent (TE) assessments will be included in the analyses. TE assessments are defined as the evaluations recorded after the first dose of study drug (AndroGel or matching placebo) but no more than 30 days after the last dose of study drug. All statistical comparisons will be performed at a 2-sided 5% significance level and confidence intervals will be constructed with 2-sided 95% confidence level. Type I error adjustment for multiple comparisons are not planned for any subgroup, supportive or sensitivity analyses.

For the primary analysis, time to events endpoints (e.g., time to MACE, time to expanded MACE) will be derived as follows: For subjects who experience the event of interest, time to event will be defined as the time from randomization to the first occurrence of the event. For a subject who does not experience an event on the study, time will be right censored at the time of his last available adequate follow-up observation that rules out the occurrence of an event. In determining the earliest occurrence of a composite or recurrent event, all incidences of that event as adjudicated by the CEC within the period of interest of the analysis will be considered.

Time-to-events of data will be summarized by number of events observed, number of subjects censored and median time to event. Median time to event (TTE) and its 95% CI,

as well as Kaplan-Meier (KM) estimates of the incidence function (cumulative event rates over time) will be calculated and plotted. A summary of continuous safety endpoints (e.g., change from baseline values in laboratory values and vital signs parameters) will include the mean, standard deviation, median and range. Categorical safety endpoints (e.g., incidence of AEs, incidences of heart failure events, incidence of prolonged QTcF or clinically significant ECG values) will be summarized using frequencies and percentages. All the summaries will be reported by treatment group and prior CV disease/risk status.

All the analyses will be adjusted for prior CV disease/risk status.

## 6.2 Analysis of Time to MACE

### 6.2.1 Primary Analysis

The primary analysis will be carried out after accrual of 256 MACE events.
Non-inferiority of TRT to Placebo in terms of risk of MACE will be evaluated by testing the following statistical hypotheses for hazard ratio (HR):

$$H_0: \text{HR} = \frac{\text{hazard of MACE on TRT}}{\text{hazard of MACE on Placebo}} \geq 1.5$$

vs.

$$H_A: \text{HR} = \frac{\text{hazard of MACE on TRT}}{\text{hazard of MACE on Placebo}} < 1.5$$

Null hypothesis ($H_0$) will be rejected with two-sided 5% level of significance if the upper limit of the 95% confidence interval (CI) for HR rule out the margin of 1.5 and the non-inferiority of TRT to placebo will be claimed.[2]

Estimation of HR

Following a Cox proportional-hazards regression model[3] the prior CV disease/risk status as a covariate will be used to estimate the HR of TRT to placebo and its two-sided 95% CI:

$$\lambda(t) = \lambda_0 \cdot \exp(\theta_1 \cdot x + \theta_2 \cdot y)$$

Where, $t$ is the time (in years) from first treatment dosing, $\lambda(t)$ indicates the hazard of experiencing MACE at time $t$, $\lambda_0$ indicates hazard at randomization, $x$ is the treatment indicator (takes value 1, if treated with TRT, otherwise 0, for placebo) and $y$ is the indicator of prior CV disease/risk status (takes value 1, if subject has history of CV disease; otherwise, 0 if subject is at CV risk but does not have history of CV disease at study entry). The two regression coefficients $\theta_1$ and $\theta_2$ will be estimated by fitting the above model to the study data. The estimated value of $\exp(\theta_1)$ represents the hazard ratio of MACE on TRT (adjusting for the effect of prior CV disease/risk status.) compared to Placebo. Effron's approximate method (1977) will be used for handling ties. In SAS, estimate of HR will be estimated using PHREG procedure and Effron's method[4] will be employed by using TIES=EFFRON option in MODEL statement.[5]

### 6.2.2 Sensitivity Analyses

The principal sensitivity analyses will be performed based on 'on-exposure' period (up to 365 days post last dose). In this analysis, MACE that occur during the period from randomization to 365 days post last dose will be included. Subjects' data will be censored after 365 days post last dose.

Two other sensitivity analyses as mentioned below will also be performed.

- The analysis only includes MACE that occur during the period from randomization to 30-days-post last dose. For subjects with events occurring after 30 days post last dose, the follow-up time will be censored at 30 days from last dose.
- For subjects with drug interruption(s) that was 3 months or longer and events occurring after 30-days-post the start date of the first interruption of 3 months or longer, the follow-up time will be censored at 30 days from first dose interruption of 3 months. For all other subjects with events occurring after 30 days post last dose, the follow-up time will be censored at 30 days from last dose.

Additional sensitivity analyses may be performed where appropriate. If pre-planned, these additional analyses will be specified in AbbVie's SAP Supplement (SAP-S) document. For all sensitivity analyses, the hazard ratio will be estimated along with its 95% confidence interval after adjusting for prior CV disease/risk.

### 6.2.3 Supportive Analysis

The absolute risk difference (RD) and restricted mean survival time (RMST) at 3-years based on the KM estimate will be calculated and summarized as continuous measures. The RMST for this CV outcome trial may be described as the MACE free expectancy over the restricted period between randomization and 3 years. RMST is defined as below:

$$\text{RMST}(3 \text{ years}) = \mu(3 \text{ years}) = \int_0^3 \Pr(T > t)\, dt$$

Under constant hazard over time assumption,

$$\text{RMST}(3 \text{ years}) = \mu(3 \text{ years}) = \frac{1 - \exp(-\lambda \cdot 3)}{\lambda}$$

where $\lambda$ is the hazard rate. Further, under constant hazard rate assumption (i.e., hazard in both control and treatment arm are $\lambda_0$ and $\lambda_1$, respectively at any given time), the ratio of RMST can be expressed as

$$\frac{\text{RMST in Treatment (3 years)}}{\text{RMST in Plaebo (3 years)}} = \frac{1 - \exp(-\lambda_1 \cdot 3)}{1 - \exp(-\lambda_0 \cdot 3)} \times \frac{1}{HR}$$

where $HR$ represents the treatment to placebo hazard ratio. As $\lambda_0$ and $\lambda_1$ are expected to be small, ratio of RMST (treatment to placebo) becomes very close to inverse of hazard ratio. With same assumptions, we can define RD as

$$\text{RD (at 3 years)} = (1 - \lambda_1)^3 - (1 - \lambda_0)^3$$

Further, with $\lambda_0 = 0.015$ (corresponding to annual event rate in Placebo arm as 1.5%) and HR = 1.5, we have the RD and difference in RMST at the end of 3rd year as follows:

$$\text{Risk difference (at 3 years)} = (1 - 0.015 \cdot 1.5)^3 - (1 - 0.015)^3 = 0.023$$

RMST in Treatment (36 months) – RMST in Placebo (36 months)

$$= \frac{1 - \exp(-0.015 \cdot 1.5 \cdot 3)}{0.015 \cdot 1.5} - \frac{1 - \exp(-0.015 \cdot 3)}{0.015}$$

$$= -0.03466 \text{ years} = -12.65 \text{ days}$$

Therefore, assuming a background event rate of 1.5% for MACE in the Placebo arm, the decision threshold (non-inferiority margin) for RD and RMST at 3 years corresponding to a HR of 1.5, is set to be 2.14% and –12 days, respectively. RMST can be calculated in SAS with TIMELIM option in PROC LIFETEST.

A comparison of the primary endpoint using RD and RMST at other time-points of interest (e.g., at 4-years) may be performed. Non-inferiority margins for these comparisons will be obtained using the same method as explained above assuming the background annualized event rate of 1.5%.

## 6.3 Analysis of Secondary CV Safety Endpoint

Time to secondary CV composite endpoint will be analyzed using similar methods as described in Section 6.2.1. A comparison of these endpoints using RD and RMST may also be performed. Non-inferiority margins for these comparisons will be obtained using the same method as mentioned in Section 6.2.3 based on the background annualized event rate of these events. These background rates and thresholds will be specified in the SAP-S document if planned.

## 6.4 Analysis of Individual Adverse Cardiovascular Events

Each of the component adverse cardiovascular events (i.e., CV death, non-fatal MI, non-fatal stroke, coronary revascularization procedures/cardiac PCI and CABG) will be analyzed separately using similar methods as described in Section 6.2.1. For each component event, deaths due to CV or any other treatment related reason will be

considered as an event. HR along with 95% confidence interval will be reported for treatment comparison.

Competing risk analysis

Two separate competing risk analyses[6] will be carried out – one for non-fatal MI and another for non-fatal stroke. In both the cases, CV death and coronary re-vascularization procedures/ cardiac PCI and CABG will be considered as competing risks.

The cumulative incidence functions (CIF) for non-fatal MI and non-fatal stroke, representing probability of failing due to these events, respectively, will be calculated as follows:

$$\text{CIF}_{\text{Non-fatal MI}}(t) = P(T \leq t, \text{ Event} = \text{Non} - \text{fatal MI})$$

$$\text{CIF}_{\text{Non-fatal stroke}}(t) = P(T \leq t, \text{ Event} = \text{Non} - \text{fatal stroke})$$

CIFs will be plotted for each treatment separately. No statistical comparison between the treatment arms will be made. In SAS, the CIFs will be produced from PHREG procedure specifying treatment as the only covariate in the MODEL statement.

## 6.5 Analysis of Tertiary Endpoints

The following endpoints will be summarized by treatment group and the prior CV disease/risk status: All-cause mortality, Heart failure events (hospitalization or urgent visit), Thromboembolic Events to include DVT/PE/venous thromboembolism (excluding superficial thrombophlebitis and Peripheral arterial revascularization).

The analyses in this section will be carried out primarily for all the events recorded.

## 6.6 Other Safety Endpoints

### 6.6.1 Analysis of Adverse Events

AEs will not be collected in the study unless they meet the definition of an AE of special interest (AESI), they resulted in study drug discontinuation or they met regulatory criteria for serious AEs (SAEs). List of AESI are included in Protocol Appendix D.

Only the above described AEs will be analyzed. Treatment emergent adverse events (TEAEs) will be analyzed separately. TEAEs are defined as any adverse event with onset or increase in severity after the first dose of study drug (AndroGel or matching placebo) and no more than 30 days after the last dose of study drug. AEs where the onset date is the same as the study drug start date are assumed to be treatment-emergent, unless the study drug start time and the AE start time are collected and the AE start time is prior to the study drug start time. If an incomplete onset date was collected for an AE, the AE will be assumed to be treatment-emergent unless there is evidence that confirms that the AE was not treatment-emergent (e.g., the AE end date was prior to the date of the first dose of study drug).

Adverse events will be coded according to the Medical Dictionary for Regulatory Activities (MedDRA) version 19.0 or higher. The Investigator will grade each AE into mild, moderate or severe based on its severity (see Study Protocol Section 6.1.2). The Investigator will also assess the relationship of the each AE to the use of study drug (see Study Protocol Section 6.1.3).

A summary of the following AEs will be provided:

- Treatment emergent AESI, regardless their relationship to study drug
- AEs leading to study drug discontinuation, regardless of their relationship to study drug
- Treatment emergent SAEs, regardless of their relationship to study drug
- Treatment emergent AESIs, possibly related to study drug
- AEs leading to study drug discontinuation, possibly related to study drug

- Treatment emergent SAEs, possibly related to study drug
- Treatment emergent AESIs, regardless of their relationship to study drug, graded as severe
- Treatment emergent SAEs, regardless of their relationship to study drug, graded as severe
- AEs leading to study drug discontinuation, regardless of their relationship to study drug, graded as severe
- Treatment emergent AESIs, possibly related to study drug, graded as severe
- Treatment emergent SAEs, possibly related to study drug, graded as severe
- AEs leading to study drug discontinuation, possibly related to study drug, graded as severe

AEs will be summarized by treatment group and overall in descending order of overall frequency by preferred term (PT), as well as in a lexicographic order system organ class (SOC) and PT. These AEs will also be summarized according to their relationship to study drug and their maximum severity. Each of these summaries will also include the number of subjects experiencing these events counting.

Deaths will be summarized using number and percentage of subjects 1) for deaths occurring while the subject was still receiving study drug; 2) for deaths occurring off-treatment within 30 days after the last dose of study drug; and 3) for all deaths in this study regardless of the number of days after the last dose of study drug.

### 6.6.2 Analysis of Laboratory Data

The lab parameters listed in Table 1 are planned to be collected for each subjects at baseline and scheduled post-baseline visits. The most recent clinical laboratory test values obtained prior to the first dose of study drug will serve as the Baseline laboratory test values.

**Table 1. Clinical Laboratory Test**

| **Hematology** | **Clinical Chemistry** | **Urinalysis** |
|---|---|---|
| Hematocrit (Hct)<br>Hemoglobin (Hgb)<br>Red Blood Cell (RBC) count<br>White Blood Cell (WBC) count<br>Platelet count (estimate not acceptable)<br>Mean Corpuscular Volume (MCV)<br>Mean Corpuscular Hemoglobin (MCH)<br>Mean Corpuscular Hemoglobin Concentration (MCHC)<br>Neutrophils<br>Lymphocytes<br>Monocytes<br>Eosinophils<br>Basophils | Blood Urea Nitrogen (BUN)<br>Creatinine<br>Total bilirubin<br>Albumin<br>Aspartate aminotransferase (AST)<br>Alanine aminotransferase (ALT)<br>Alkaline phosphatase<br>Gamma-glutamyl transferase (GGT)<br>Sodium<br>Potassium<br>Calcium<br>Inorganic phosphate<br>Uric acid<br>Cholesterol<br>LDL-C<br>HDL-C<br>Total protein<br>Glucose<br>Triglycerides<br>Bicarbonate/$CO_2$<br>Chloride | Specific gravity<br>Ketones<br>pH<br>Protein<br>Glucose<br>Blood<br>Leukocytes<br>Nitrites<br>**Other Laboratory Tests**<br>hsCRP<br>PSA<br>Serum Creatinine<br>Fasting Plasma Glucose<br>HbA1c<br>**Sex Steroids**<br>Testosterone<br>Free Testosterone<br>Dihydrotestosterone (DHT)<br>Estradiol |

If more than one measurement exists for a subject on a particular day, an arithmetic average will be calculated. The average will be considered to be the subject's measurement of that day.

For each of the continuous laboratory parameters, the following outputs will be produced:

- Summary of lab measurements at each scheduled visits
- Summary of changes from baseline at each post-baseline visits along with treatment group difference
- Shifts from baseline categories to worst and final post-baseline categories in terms of normal ranges. Categories can be lower limit of normal (LLN), normal and upper limit of normal (ULN).

In addition, the incidence of the following lab abnormalities/events will also be summarized:

- Confirmed increase > 1.4 ng/mL in prostate specific antigen (PSA) above baseline during the first year [> 0.7 ng/mL in men on 5-Alpha Reductase Inhibitor (5-ARI)]
- Detection of a new prostate nodule or induration
- Confirmed absolute PSA value > 4.0 ng/mL at any time during the study (> 2.0 ng/mL in men on 5-ARI)
- Men 45 – 54 years of age whose Baseline PSA was < 1.5 ng/mL and whose PSA increases to > 3.0 ng/mL at any time during the study
- Confirmed Hct > 54% while subject is at lowest dose (20.25 mg/dL)
- Total serum testosterone > 750 ng/dL while subject is at lowest dose (20.25 mg/dL)

Available treatment emergent (TE) lab values values at the time of reporting will be included in the analyses. Unscheduled assessment will not be included in the summary of change from baseline, but will be included in producing shift tables and summary of lab abnormalities.

### 6.6.3 Analysis of Vital Signs and Weight

Vital sign parameters such as SBP, DBP, pulse rate, temperature, BMI and weight are collected at baseline and scheduled post-baseline visits. The most recent values obtained prior to the first dose of study drug will serve as the Baseline values. If more than one measurement exists for a subject on a particular day, an arithmetic average will be calculated. The average will be considered to be the subject's measurement of that day.

For each of the vital signs parameters, the following outputs will be produced:

- Summary of lab measurements at each scheduled visit
- Summary of changes from baseline at each post-baseline visit along with treatment group difference

Available treatment emergent (TE) vital signs values at the time of reporting will be included in the analyses.

### 6.6.4 Analysis of Electrocardiogram (ECG) Parameters

A resting 12-lead ECG will be performed at the designated study visits as specified in Protocol Appendix C and QT interval corrected using Fridericia's formula (QTcF) will also be determined. Clinical significance of any abnormal finding will be interpreted by a qualified physician. Only incidence of prolonged QTcF (i.e., QTcF > 430 msec) and clinically significant findings will be reported in eCRF and summarized. Following summaries will be provided:

- Proportion of subjects with QTcF > 430 msec
- Proportion of subjects with clinically significant findings
- Proportion of subjects with QTcF > 430 msec and/or clinically significant findings

Available treatment emergent (TE) values at the time of reporting will be included in the analyses.

## 7.0 Efficacy Analyses

Efficacy endpoints from the sub-studies will not be covered in this main SAP. Separate SAP supplements will accompany this main SAP at a later date describing the full analyses for the efficacy sub-studies.

## 8.0 Summary of Changes

### 8.1 Summary of Changes Between the Previous Version and the Current Version

Not applicable.

### 8.2 Summary of Changes in Previous Version

Not applicable

## 9.0 References

1. Study protocol of M16-100, Amendment 1. 26 February 2018.
2. FDA guidance on non-inferiority clinical trials to establish effectiveness. November 2016.
3. Cox DR. Regression models and life tables (with discussion). J R Stat Soc Series B. 1972;20:187-220.
4. Efron B. The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc. 1977;72:557-65.
5. SAS Institute Inc. SAS/STAT® 13.1 User's Guide. Cary, NC: SAS Institute Inc.; 2013.
6. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. 2nd ed. New York: Springer-Verlag; 2003.