The **G**lycemia **R**eduction **A**pproaches in **D**iabetes: A Comparative **E**ffectiveness **Study (GRADE Study)**

NCT01794143

Statistical Analysis Plan

October 5, 2017

Sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

GRADE Study Coordinating Center
Biostatistics Center
George Washington University
6110 Executive Boulevard
Rockville, Maryland 20852

07/11/2017
Revised 10/05/2017

The **G**lycemia **R**eduction **A**pproaches in **D**iabetes: A Comparative **E**ffectiveness **Study**
(GRADE Study)

STATISTICAL ANALYSIS PLAN

This preliminary Statistical Analysis Plan was prepared as a supplement to the U01 renewal application for the second 5 years of funding for GRADE by the George Washington University Biostatistics Center. The SAP will be further refined and then "locked" prior to the completion of follow-up in the third quarter of 2021.

1 INTRODUCTION AND BACKGROUND

1.1 Introduction

This Statistical Analysis Plan (SAP) is being submitted as an attachment to our application for the Biostatistics Research Center (BRC) that is being submitted in response to RFA-DK-16-511. The overarching goal of this application is to continue and complete the GRADE study.

The SAP describes general statistical considerations and strategies that will be used to address specific objectives and aims, the statistical methods to be employed in specific types of analyses, the statistical analyses to be employed to assess each specific study outcome, and finally other statistically related activities of the BRC.

This document supersedes and provides additional details beyond those specified in the original protocol and may specify new or improved methods for specific analyses based on the latest developments in statistical methodology. For example, the closed testing procedure of Marcus, et al. (1976) will be employed in lieu of the Holm (1989) procedure specified in the original protocol owing to simulation studies that we have conducted showing that the former has greater power. The primary and secondary outcomes and analyses are stated in the protocol synopsis.

This document will be refined as analysis plans for the major study papers are developed by writing committees to be designated by the study leadership. The analysis plan will be formally "locked" prior to data lock.

1.2 Study Population

The objective of GRADE is to assess the long-term beneficial and adverse effects of four commonly used treatments for Type 2 Diabetes Mellitus (T2DM) in a cohort of up to 5000 subjects who are generally representative of the general US population, including African Americans, Hispanic Americans and other minorities. A total of 36 sites (some with sub-sites) were selected to recruit subjects from their local populations. Table 1 presents the baseline characteristics of the 4001 patients randomized into GRADE as of November 13, 2016. These baseline characteristics reflect the eligibility criteria for enrollment and the natural history of diabetes. The mean age is 57 and 40% are at least 60 years of age. Sixty-three % are men (the relatively large fraction of men reflects the recruitment from 10 VA clinical sites). The majority (65%) are white with 20% African-American and 18% Hispanic/Latino. The mean diabetes duration is 4 years. At the time of randomization, after the titration of metformin during the run-in period, the mean metformin dose is 1947 mg/day. The mean weight is 100 kg with systolic/diastolic blood pressures of 128/77 mmHg. The mean HbA1c is 7.5% with 85% being at least 7%.

1.3 Study Performance

Table 2 presents metrics to describe the performance of the study including the completeness of data collection, compliance with specific study assessments, and study follow-up through October 31, 2016. Of the 9520 who attended an initial screening visit 3957 (41.6%) were randomized, not including participants who have entered the screening process but whose randomization is pending. Over an average of 16 months of follow-up, 20 patients withdrew from continued follow-up in the study (i.e. withdrew consent) and 15 died.

Table 1. Baseline Characteristics of the GRADE Cohort Recruited as of November 13, 2016.	
Data are mean ± SD or n (%)	**Total**
N	4001
Age (years)	56.9 ± 9.9
>= 60 yr	1618 (40.4%)
Gender (% males)	2520 (63.0%)
Race	
American Indian / Alaska Native	116 (2.9%)
Asian	139 (3.5%)
Native Hawaiian or Other Pacific Islander	22 (0.5%)
Black or African-American	811 (20.3%)
White	2609 (65.2%)
More than one race/other	250 (6.2%)
Unknown or not reported	54 (1.3%)
Ethnicity (%)	
Hispanic/Latino	733 (18.3%)
Not Hispanic/Latino	3238 (80.9%)
Unknown	30 (0.7%)
Age at diagnosis (yr)	52.8 ± 9.6
Duration of diabetes (yr)	4.0 ± 2.7
Current Metformin dose mg/day)	1947.4 ± 197.9
Weight (kg)	100.1 ± 22.5
Diastolic BP (mmHg)	77.2 ± 9.8
Systolic BP (mmHg)	128.1 ± 14.7
HbA1c (%)	7.5 ± 0.5
>= 7%	3404 (85.1%)
Laboratory Tests	
Serum Creatinine (mg/dL)	0.8 ± 0.2
Cholesterol (mg/dL)	163.7 ± 37.3
Triglycerides (mg/dL)	154.5 ± 125.0
High Density Lipoprotein (HDL, mg/dL)	43.3 ± 12.2
Low Density Lipoprotein (LDL, mg/dL)	90.5 ± 31.0
Urine albumin:creatinine ratio (ACR) (mg/g)	6.6 (3.2, 17.4)
eGFR (ml/min/1.73m²)	95.4 ± 17.0
Data are mean ± SD, median (IQR), or n (%)	

Of all the visits expected, 95% of quarterly and 97% of annual visits were conducted. For all other items, the number of procedures, collections or measurements exceeds 90% of the number expected had there been complete (100%) compliance with the protocol, e.g. the number obtained relative to the number expected among those surviving.

Table 2. Study Performance Metrics as of October 31, 2016	
N and % of the number expected (where available)	**Total**
Number of Participants Screened	9520
Number of Participants Randomized	3957
Average Duration of Participant Follow-up (months)	16.5
Number of Participants Withdrawn from Study	20
Study Visits Completed	
Number of Quarterly Visits:	15561 (95.2%)
Number of Annual Visits (1-3):	3447 (97.0%)
Year 1	2393 (97.0%)
Year 2	962 (96.9%)
Year 3	92 (100.0%)
Number of Assessments Completed	
ECG	4842 (97.8%)
Neuropathy (MNSI)	7238 (96.4%)
Neurocognitive	3939 (99.5%)
Oral glucose tolerance test (OGTT)	6323
Physical Assessments	
Blood Pressure	7237 (95.6%)
Weight	7228 (95.4%)
Quality of Life	
QWB-SA	7242 (96.4%)
SF-36	7245 (96.5%)
Laboratory Tests	
HbA1c	23024 (95.4%)
Urine albumin:creatinine ratio (ACR)	8434 (92.6%)
Serum creatinine/eGFR	7396 (97.7%)
Fasting Lipids	7211 (95.2%)
Number of DNA Collections	3688 (93.2%)
Number of Stored Sample Collections	6902 (93.4%)
Number of SAEs Reported	640

2 STATISTICAL CONSIDERATIONS

2.1 Analysis Data Sets (Intention-to-Treat, Per Protocol)

GRADE has been designed, and is being implemented, under an intent-to-treat design whereby all subjects randomized are included in the study for the study duration, and asked to continue follow-up and outcome assessments, regardless of other outcomes such as compliance with the assigned medications, termination of treatment for adverse effects or missed visits. Thus, unless specified otherwise, analyses will first be conducted using the ***intent-to-treat data set*** that includes all subjects randomized into the study, including those who did not receive a dose of the assigned study medication, and all observed data from each subject regardless of compliance and adherence with the assigned treatment regimen or with the follow-up schedule. Where specified, additional analyses may also be based on the ***modified intent-to-treat data set,*** excluding patients who did not receive the assigned study drug, for whatever reason.

2.2 Treatment Group Comparisons and Adjustment for Multiple Tests.

The original protocol specified that multiple tests among groups would employ the Holm improved Bonferroni procedure (Holm, 1989) to adjust the levels of significance required to protect against inflation in the type I (false positive) error probability in the set of 6 pair-wise comparisons among the four study groups. However, by simulation we have shown that the closed testing procedure (Marcus et al, 1976, Chi, 1998) is more powerful than the Holm procedure.

Under the closed testing approach, a sequence of tests is conducted. First the set of four treatment groups will be compared with an omnibus T^2-like test for any difference among the 4 groups. If that test is significant at the specified significance level (α=0.05), then under the closed testing procedure, each of the additional 3 group sub-hypotheses are then tested at the same level α. Then if any two of these sub-hypotheses are rejected at level α, the common two group components can be tested also at level α. For example, if the 3 group test of equality of groups 1, 2, 3 and 1, 2, 4 are both significant at level α, then equality of groups 1 and 2 can be tested at that level α. Note that all tests are conducted using an uncorrected α=0.05.

The primary interest in the above analyses is the comparison of each pair of groups so as to determine which treatments are better in relation to a specific alternate treatment. An additional question of interest is whether a given treatment is on average superior to the other treatments in combination. This would entail 4 tests, each using the full cohort, group 1 versus 2, 3, 4; group 2 versus 1, 3, 5, group 3 versus 1, 2, 4 and group 4 versus 1, 2, 3. Again, rather than impose an alpha penalty of 0.05/4 for multiple tests, these analyses can be conducted under the closed testing principle starting with the overall test of the difference between the 4 treatment groups.

These analysis strategies will be employed to compare the treatment groups for all study outcomes.

2.3 Prevalence Analyses (Binary Outcomes)

Examples of such a binary outcome include the presence or absence of macroalbuminuria at each year of follow-up. Such analyses of a binary variable typically describe the *prevalence* of an outcome at a specific point in time. Logistic regression models (Lachin, 2011) will be employed to examine the effects of factors (e.g. treatment group) on the odds of the binary outcome at that time (the odds ratio). The analysis can also use other covariates in the model to adjust for the effects of the covariates on the outcome, and can use an interaction between the

main factor and the covariates to assess the homogeneity of the factor effect over levels of the covariate(s).

In these models, likelihood ratio tests of effects will be employed and the strength of the effect measured by a partial entropy R^2 for each covariate (Lachin, 2011). Value-added plots (Pregibon, 1981) will be employed to explore whether transformations or polynomial covariate effects are warranted rather than a simple linear effect. Goodness of fit will be assessed by the Hosmer-Lemeshow test and over-dispersion using the tolerance limits on the ratio of the Pearson Chi-square to its df (Lachin, 2011). If the model assumptions are violated, the robust estimate of the covariance matrix of the estimates will be employed as the basis for confidence intervals and tests of significance (Lachin, 2011).

Generalized estimating equations (Diggle et al, 2002) with a logit link will be employed to assess the effects of covariates on the odds of an outcome over repeated points in time, allowing for the correlation among the repeated measures. Partial Wald or score tests will be used to test covariate effects and Madalla's R^2 (Lachin, 2011) used to describe the strength of effect for each covariate.

2.4 Cumulative Incidence (Life-Table) Analyses

A principal set of outcome analyses will consist of survival (life-table) analyses of time-to-event outcomes such as the primary metabolic outcome.

2.4.1 **Continuous Time Observations**.

Event times are obtained in continuous time when the day or date of an event is known, and the date at which the subject was last at risk (the right censoring time) is known. Examples are the times of death or myocardial infarction, etc. Analyses of such data will be performed using the standard Kaplan-Meier estimate of the survival or cumulative incidence function. The unadjusted log-rank test will be used to test for differences between treatment groups (Kalbfleisch, Prentice, 2002, Fleming, Harrington, 1991, Lachin, 2011). Analyses would be conducted to compare treatment groups adjusting for baseline characteristics if there are concerns for confounding or imbalances, or to improve power. The proportional hazards regression model (Kalbfleisch, Prentice, 2002, Cox, 1972, Martinussen, Scheike, 1991) would be employed to adjust for a set of covariates, or to jointly assess the influence of a set of factors simultaneously.

If tests of the proportional hazards assumption do not apply, inferences (confidence intervals and p-values) will be obtained using the robust information sandwich estimates of standard errors.

In exploratory analyses, the assumption of proportionality will be tested using the test of Lin (1991) and using other graphical methods (Therneau, Grambsch, 2000). If non-proportionality is found, then either an alternate model may be employed, such as the proportional odds model (Bennett, 1983, Younes, Lachin, 1997) or transformations of the covariates may be employed, or time effects may be included in the model. Alternatively, since the coefficient estimate under non-proportional hazards still converges to a finite constant, this could be interpreted (approximately) as an average log hazard ratio and the precision (SE) and significance assessed by the robust covariance estimate of Lin and Wei (1978), and the robust model score test will be used to assess group differences (Lachin, 2011).

2.4.2 **Grouped Time**

In many instances, however, the exact time of an event is not known, such as when the primary metabolic outcome is first observed from an HbA1c value at a quarterly visit (subsequently confirmed) and we only know that the "event" may have occurred any time between the current and last evaluation. For outcomes observed with a fixed schedule over time, since all subjects

have the same schedule of assessments (e.g. eGFR annually), a fairly standard simple procedure can be employed. Basically, for analysis of annual renal assessments, the time to a renal event (e.g., CKD3) employed in the analysis is the scheduled time of the evaluation in whole years (1, 2,...) rather than the exact study day or fractional year of the visit. Patients who remain event-free will have a right censored time (period of observation) as of the day last evaluated. Since the outcome can only be observed when an examination is conducted, this leads to the construction of a modified Kaplan-Meier survival (or cumulative incidence) function (Lachin, 2011). In a proportional hazards analysis, the discrete logistic model of Cox (1972) will be employed. With frequent monitoring, Lachin (2013) shows that this discrete time analysis provides nearly the same level of power as would an analysis where the actual event is observed in continuous time.

Poisson regression models (Lachin, 2011,McCullagh, Nelder, 1989) may also be applied to such discrete interval data (Laird, Oliver, 1981, Whitehead, 1980). These models have the advantage of modeling the absolute risk rather than the relative risk as is the case for the proportional hazards model. This model also readily allows use of time-dependent covariates. These models require that one either assume that the background hazard is constant over time or that it can be modeled by covariate effects in the model. The proportional hazards model, however, conditions on the variation in the background hazard function so that it is not explicitly estimated as part of the model.

2.4.3 **Confirmed Outcomes**

Some outcomes evaluated in grouped time will require that the event be confirmed on two successive measurements, such as the primary metabolic outcome that requires confirmation of an initial HbA1c ≥ 7% (the "triggering" value) at the next quarterly visit (the "confirmation" value), or sooner if the triggering value is > 9%. In this case it is possible that at the last visit of a subject, his/her HbA1c meets the criterion for a triggering value but there is no opportunity to obtain a confirmation value. In this case the primary outcome status is unknown. Therefore, two possible scenarios should be carefully considered to ensure that all participants' censoring times are correctly defined.

(i) The event time of participant *i* is considered *right censored* at the final quarterly visit, say b_i, if HbA1c is <7%, and the confirmed primary outcome has not been reached throughout the study.

(ii) The event time of a participant will be considered *right censored* if HbA1c is $\geq$7% at the final quarterly visit, and HbA1c is < 7% at the previous quarterly visit, say a_i. In this case, a follow-up confirmation is not possible. The event time for subject *i* will be considered right censored at time a_i.

Because the assessments are performed quarterly, the event time will be the discrete quarterly follow-up visit number at which HbA1c is $\geq$7%, subsequently confirmed. When examinations are performed frequently, Lachin (2013) shows the gain in efficiency is negligible when one considers the time to event as interval censored data. Therefore, it is reasonable to consider the event times as right censored observations and to ignore the time intervals between assessments.

2.4.4 **Interval Censored Observations**

However, the simple grouped time methods above would not apply to the analysis of an outcome with widely varying intervals between examinations, which may apply to the GRADE metabolic outcomes, especially the tertiary outcome (requiring initiation of an intensive basal/bolus insulin regimen). Such data are *interval censored* because only the interval of time in which an even occurred is known, and the intervals may differ among patients. For interval-censored event time data, methods are also available that take into account the exact day of each successive visit and the length of the exact interval in days between successive visits.

Turnbull (1976) described an estimator of the survival distribution (event-free distribution) for such interval-censored data and Finkelstein (1986) described a generalization of the proportional hazards regression model to such data. However, both procedures require the estimation of a large number of nuisance parameters to describe the underlying background survival distribution. Younes and Lachin (1997) described a family of regression models that provide a regression spline estimate of the background hazard (and thus cumulative incidence) functions and that include the proportional hazards and proportional odds models as special cases. Therefore, this procedure also provides a generalization of the log-rank test to such data. See also Pan (1999), Boruvka and Cook (2015) and Wang et al. (2016).

These methods are non-parametric in that no form of the underlying hazard function is assumed. However, they involve various nuisance parameters that must also be estimated to fit the model. Another approach is to employ a parametric model with a specific underlying hazard function with only one extra shape parameter, such as an accelerated failure time model using the SAS PROC LIFEREG to describe covariate effects on the time acceleration factor (2002). Such models, however, are not directly interpretable in terms of the covariate effects on the underlying hazard or survival functions. Rather, a parametric model, such as the Weibull model of Odell, et al. (1992) could be employed that yields an estimate of the covariate effects on the relative risk of the event over time, in the same manner as the expression of covariate effects in the Cox PH model. The model can be fit using a Weibull accelerated failure time model from which the Weibull model parameter estimates and covariance matrix can be obtained (Lachin, 2011). Weibull model analyses that employ fixed and/or time-dependent covariates can also be obtained from the models of Sparling, et al. (2006). This model was used to assess time-dependent covariate effects on the risk of retinopathy progression during EDIC (DCCT/EDIC Research Group, 2015).

2.4.5 **Competing Risks**

The risk of some events will be curtailed due to competing risks, such as the analysis of the incidence of a cardiovascular event where some subjects die before such an event occurs. In this case, the deaths are not simply right-censored. Nevertheless, a Cox PH model analysis of the event time with right censoring on death still has a valid interpretation as the effect of the model covariates on the cause-specific hazard function for the event (Prentice et al, 1981).

A more precise analysis would be to describe a true estimate of the cumulative incidence of the index event (e.g. laser therapy for retinopathy) adjusting for the incidence of the competing risk (mortality), such as an estimate of the sub-distribution function for the index event (Gray, 1988, Pepe, 1991, Pepe and Mori, 1993). Fine and Gray (1999) also provide an extension of the Cox PH model to the analysis of covariate effects on the cumulative incidence function itself that accounts for covariate effects on both the cause specific hazard function for both the index and competing risk events. These approaches are especially useful when there are differences between groups in the incidence of mortality itself which must be considered in addition to the differences in the incidence of the outcome (e.g. CVD event).

The methods for competing risks extend in a similar fashion to applications in which a study subject can move among a number of $k>1$ states over the course of the study, called multistate models (Kalbfleisch, Prentice, 2002, Andersen et al, 1993, Beyersmann et al, 2012).

2.5 Incidence of Recurrent Events

In some cases, a subject may experience the same or like events over time, such as recurrent hospitalizations. Most such recurrent event outcomes will be observed in calendar (continuous) time. For such data, Andersen et al. (1982) describe methods for the estimation of the underlying incidence rate function over time and develop a generalization of the logrank and other tests of significance of differences between groups with respect to the incidence function over time (Lachin,

2011). The incidence rate (intensity) function estimates can be smoothed using a kernel-smoothed estimator as described by Ramlau-Hansen (1983, 1983). To account for the effects of covariates on the incidence rate, either the Poisson regression model (Lachin, 2011, McCullagh, Nelder, 1989) or the multiplicative intensity model (Lachin, 2011,Fleming, Harrington, 1991, Andersen et al, 1993, Andersen, Gill, 1982) will be employed. The multiplicative intensity model is a generalization of the proportional hazards model which allows for recurrent events in the same subject over time. However, it does so using a rather unrealistic assumption that the successive event times are conditionally independent of those that preceded. This assumption was relaxed in the proportional rate model of Lin et al. (2000) that also employs the robust information sandwich estimate of the covariance matrix of the coefficient estimates. These models can also be employed to assess the association between outcomes and a time-dependent covariate. These methods can be employed for the assessment of the differences between treatment groups in the risk of hypoglycemia.

2.6 Rates of Events

In some cases, however, such as episodes of hypoglycemia, the exact dates of recurrent events may not be known. Rather, only the number of such events over an interval of time is reported. The incidence of such events will be summarized as a crude rate. Such rates will be presented as the number of events per 100 patient-years based on the ratio of the observed number of events to the total patient-years of exposure. The standard error for such rates will be computed allowing for "over-dispersion," i.e. assuming that the subjects have some underlying distribution of intensities (hazards) rather than the usual restrictive assumption that the same intensity applies to all subjects (Lachin, 2011). The risk ratio (relative risk) will be used to summarize the difference between groups, and tests will be based on the large sample estimate of the variance of the log relative risk.

Poisson regression models will be employed to assess covariate effects on the rate of such events (Lachin, 2011), expressed as a risk ratio (relative risk), and robust methods for inference will be employed if the model Poisson assumptions are violated (Lachin, 2011). If a preliminary test of the homoscedastic Poisson assumption is significant, then either a zeros inflated Poisson model or alternate parametric models such as a negative binomial model will be employed (Lachin, 2011). With longitudinal observations, we will consider models allowing the underlying baseline intensity to change with time using nonparametric tests (Thall, Lachin, 1988) and mixed or marginal Poisson models (Lawless, Zhan, 1998, Chen et al, 2005).

2.7 Ordinal Outcomes

An ordinal outcome is a nominal assessment with multiple (>2) categories with an implied ordering, such as no nephropathy, microalbuminuria only, albuminuria only, or end-stage renal disease at a point in time. Simple proportions in each category will be used to describe the prevalence within each category at a given point in time, and differences between groups tested using the 1 df Mantel-Haenszel test of mean scores (Agresti, 1990), or using the Wilcoxon signed rank test with the adjustment for tied ranks (Snedecor, Cochran, 1980). A proportional odds model (Agresti, 1990) will be used to examine covariate effects on the prevalence within each ordered category. If the test of the proportional odds assumption is rejected, then that implies the need to model covariate effects on each category separately. In this case, the odds of each category versus a designated reference category (e.g. no nephropathy) at a specific point in time will be assessed using a multinomial logit model (Agresti, 1990). In essence, this model simultaneously fits a logistic model for C-1 comparisons of each positive category versus the reference category. The results of these models will be summarized as above for a logistic regression model. For a longitudinal analysis of covariate effects on repeated ordinal assessments over time, a proportional odds model with generalized estimating equations (GEE)

will be employed (Parsons et al, 2009). Alternately, the difference between groups in the longitudinal ordinal assessments can be tested using the Wei and Lachin (1984) multivariate rank test.

2.8 Analyses of Quantitative Data

For quantitative (numerical) variables with no point of truncation, e.g. the albumin-creatinine ratio in mg/g, simple differences between groups will be assessed by a Wilcoxon test (Snedecor, Cochran, 1980). Models adjusting for covariate effects will be conducted using normal errors regression models (Neter et al, 1996). Partial residual or value-added plots will be employed to determine whether a transformation or a polynomial best represents a covariate effect rather than a simple linear term. The homoscedastic normal errors assumptions will be tested using the Shapiro-Wilks test of normality of residuals and White's test of homoscedasticity of error variances (1980). If violations are detected, then an appropriate transformation will be sought. If still violated, all inferences will be based on White's robust estimate of the covariances of the estimates (White, 1980) that provides consistent estimates of the variances of the coefficient estimates.

2.9 Marginal Repeated Measures Analyses

Many assessments are repeated at intervals during GRADE for which repeated measures analyses will be conducted. Most such analyses will employ multivariate methods for the analysis of repeated quantitative, ordinal or qualitative measures.

The normal errors mixed model will be employed for an analysis of covariate effects on repeated quantitative measures over time using an "unstructured" covariance matrix for the repeated measures (Diggle et al, 2002, Demidenko, 2006). Such "marginal" analyses provide an assessment of covariate effects on the average of values over time, or at specific points in time when covariate by time effects are employed. For example, these models will be used to evaluate the interaction between treatment group and time to determine if there were persistent treatment group differences in eGFR levels over time.

For variables that do not satisfy the normal errors assumption, or those that are ordinal or nominal in nature, alternate methods may be employed. These include the multivariate non-parametric Mann-Whitney rank analysis for quantitative or ordinal measures (Thall, Lachin, 1988, Wei, Lachin, 1984, Lachin, 1992) and the multivariate analysis of qualitative observations (Lachin, Wei, 1988). These methods are intrinsically marginal in that the treatment group difference is assessed at each point in time, and an overall assessment is derived by pooling the results over time. In the simplest case of a binary outcome variable, e.g. albuminuria present or absent, the marginal analysis consists of the comparison of the simple prevalences (proportions present) at each visit, which are then used to compute ~~a~~ risk difference (or relative risk or odds ratio) at each visit, which are averaged over all visits.

In these analyses, a variety of multivariate tests of significance can be used (Lachin, 1992). A commonly used test is based on the minimum variance efficient weighted average of the summary measures of treatment group differences over time (Mann-Whitney differences, odds ratios, etc.), termed the test of aggregate association. This test, analogous to the Mantel-Haenszel test, is appropriate when a common value of the summary measure is assumed to exist. Alternately, the Wei-Lachin test of stochastic ordering, or multivariate one-directional test, is more general in that it tests the hypothesis of no difference over time against the alternative hypothesis that the summary measures for all visits tend to differ in the same direction over time, such as where the outcome values tend to be systematically higher (or lower) in one group than the other. This test is based on the unweighted simple average of the summary measures and has been shown to be a maximum efficient robust test against the family of alternatives where the groups differ in the same direction

over time, but not to the same degree (Frick, 1994). Lachin (2014) also shows that this simple test also provides an efficient method for the assessment of treatment group differences in a set of multiple outcomes.

To evaluate the effects of covariates, including time-dependent covariates, on quantitative or qualitative outcomes over time, regression models based on the method of GEE (Diggle et al, 2002, Liang, Zeger, 1986, Zeger, Liang, 1986, Fitzmaurice et al, 2011) will also be employed. This method can be used to estimate a common covariate effect for all visits over time, or visit specific effects can be estimated which can then be used in a test of stochastic ordering if desired.

2.10 Random Effects "Growth Curve" Models

For some outcomes, the longitudinal rate of change of the outcome over time will be analyzed based on the within-subject "slopes" of the regression of the outcome on time. These are commonly known as growth curve analyses. These analyses are especially common in the analysis of measures of renal function and are commonly employed in the analyses of the rate of change in eGFR over time.

A general family of such models has been described by Laird and Ware (1982) and Jennrich and Schluchter (1986), among many others. Laird and Ware (1982) referred to the simplest form of these models as the "two-stage" random effects model. These models assume a common "shape" to the regression of the outcome over time for each subject (e.g. linear, quadratic, log-linear, etc.) with a corresponding within-subject component of variance, and then assume that the regression parameters in the population of subjects have some overall distribution with an average curve over time (e.g. mean intercept and slope) and between-subjects variance components. Usually, this "mixing" distribution is assumed to be multivariate normal. Given the assumed shape of these curves and the assumed mixing distribution, estimates of the average parameters (mean intercept and slope) and the within- and between-subjects variance components are obtained.

These models can incorporate the effects of subject-specific and time-specific covariates. Therefore, such models can be used to describe the average pattern of change in the outcome over time and to assess the effects of various covariates on the average values at any point in time, or on the pattern of change over time (Vacek et al, 1989).

These mixed models (Demidenko, 2006) are essentially parametric in that they assume that the within-subject residuals are normally distributed and that patient-slopes in the population are also normally distributed. For some measures such as the urine albumin:creatinine ratio (ACR), these assumptions may not apply. In these cases, it will be necessary to explore a transformation of the data, such as the log transformation, which improves the distributional assumptions of the model. For some measures, such as ACR, a log transformation may be more biologically meaningful. When the rate of change in an individual subject is described on the log scale, it is implied that the percentage change over time is a constant for each subject rather than the absolute magnitude of the change being a constant for each subject, as is implied by a linear slope in the original measurements.

2.11 Informatively Censored and Missing Observations

All of the above methods assume that missing values are missing at random (Little, Rubin, 2002), and (efficient) unbiased results can be obtained using the direct likelihood method or EM algorithm (Molenberghs, Kenward, 2007), multiple imputations (VanBuren, 2012, Carpenter, Kenward, 2013) and inverse probability weighting (Molenberghs et al, 2015). This assumption, however, may not be appropriate in some instances.

For point prevalence analyses, other informative mechanisms in addition to mortality may apply, such as where patients who develop congestive heart failure are unable to attend the clinic visit for other outcome assessments. These instances are more problematic because some

assumptions are then required regarding the nature of the association between the reason for missing data (termed "missingship") and the values of the missing observations. In the case where patients who have informatively missing values are assumed to have "worse" values than any observed non-missing values, then a rank analysis can be performed with a worst rank score assigned to the informatively missing observations (Lachin, 1999). For example, patients who have died are usually assumed to have a worse quality of life than that of those who survive and complete a quality of life questionnaire.

For longitudinal growth curve analyses, various methods have been proposed (Wu, Bailey, 1988, Wu, Bailey, 1989, Wu, Carroll, 1988, Wu et al, 1994, Schluchter, 1998). Some of these methods estimate the relationship between the repeated measures within subjects and the likelihood of informative censoring, which is then used to obtain a less biased estimate of the overall mean curve parameters (intercept and slope). Another approach which has been shown to be unbiased, but not as efficient under weaker assumptions, is to simply use an unweighted average of the within-patient coefficients (Wu et al, 1994).

2.12 Analyses that Adjust for Missing Data

As described in Section 1.4 above and Table 2, the level of completeness of the study data has thus far been outstanding. Nevertheless, by the end of the study, the fraction of missing data in some specific analyses may be higher than desired. In this case the amount and patterns of missing data, and its association with other variables, will be explored so that an appropriate statistical method for analysis can be employed.

Missing data fall into three categories (Little, Rubin, 2002): missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). The statistical implications differ for each category. MCAR refers to data that are missing purely by chance, i.e. for reasons that are totally random and unrelated to both the independent variables (e.g. treatment group) and the outcome variable (e.g. CVD). In this case a “complete case” analysis of the subjects with complete data can be unbiased. Unfortunately, there is no way to prove that this assumption applies, although it may be possible to show that it does not, such as when the characteristics of those with missing data differ significantly from those with complete data.

Missing at random (MAR) refers to missing data that can be related to other observed data, such as where missing data can be more prevalent among males than females, or associated with the prior patterns of changes in the outcome or other covariates. If the important factors can be measured and used as adjusting covariates in the model then the analysis results will be unbiased. In these cases, multiple imputation (Rubin, 1987), likelihood-based analysis with computations using the expectation-maximization (EM) algorithm (Dempster et al, 1977), or inverse probability weighting (Seaman, White, 2013) methods might be employed. The precise imputation schemes and interpretation of results will depend on the characterization of the missing data mechanism, as best can be determined (Little, Rubin, 2002, Rubin, 1987, Schafer, 2000).

A simple illustration is provided by a case-control design where a biomarker is evaluated in relation to an outcome from which the cases are defined. The study could employ a case-cohort subsample consisting of predefined numbers of cases and controls who were selected at random (separately) from the cases and controls in the full cohort, perhaps with different sampling probabilities. The sampling probability for the cases is the ratio of the number of cases in the sample to the number of cases in the full cohort, and likewise the probability for controls. Unbiased results can then be obtained using weights that are the inverse of the sampling probabilities for the cases and controls.

Missing not at random refers to cases where the MAR assumption does not apply. When appropriate, in the case of non-ignorable missing data (i.e., MNAR), sensitivity analyses will be

performed using pattern-mixture (Little, 1993, Little, Wang, 1996) or selection models (Hogan, Laird, 1997) to evaluate the robustness of our conclusions to a range of sensible conditions.

2.13 Covariate Adjustments, Effect Modification and Mediation

The term covariate is used to refer to any variable or factor that is included in statistical models to describe the effects of the variable or factor on an outcome, such as a model with treatment group and the baseline HbA1c as factors in a Cox PH regression model of the primary metabolic outcome. The model then provides estimates of the hazard ratios among treatment groups and the log hazard ratio per unit increase in HbA1c. Covariates may be included individually or jointly. Covariates may be fixed (e.g., gender, race-ethnicity, randomized treatment group, baseline value) or time-varying (e.g., HbA1c, lipids), depending upon the adjustment desired.

Often baseline covariates are known to be strong risk factors for an outcome, such as age and sex for analyses of CVD outcomes. Adjusting for such baseline covariates in a model that also compares treatment groups will adjust for any baseline imbalance among groups being compared and will increase power.

All analyses of post-randomization characteristics will be assessed as time-dependent covariates. Such analyses will provide a prospective interpretation of the effect of the time-dependent characteristic. For example an analysis of the effects of HbA1c over time on the risk of developing albuminuria will describe the hazard ratio of albuminuria in the future per unit increase in HbA1c at any point in time during follow-up. In such analyses the time-dependent covariate may have an acute or instantaneous effect on the outcome, or it may have a long-term chronic effect. If the former the current value of the covariate might be employed whereas for the latter the updated mean value since baseline might be employed. For example, the current HbA1c value might be used in relation to the risk of hypoglycemia (an acute effect) whereas the updated mean HbA1c might be used in relation to the risk of albuminuria.

Models may also test for an interaction between a baseline covariate and group to determine whether there is statistically significant heterogeneity (or interaction) in the group effect over levels of the covariate, i.e. whether the covariate is an ***effect modifier***. In this case the treatment group difference is described within categories or levels of the adjusting covariate. If significant heterogeneity or interaction is detected among subgroups at the 0.05 level, then the closure principle (Marcus et al, 1976) can be employed to test the difference between groups. If there are only two subgroups then the group effect within each can be tested at the 0.05 level without the need to adjust for multiple tests.

Additional analyses may then be conducted to assess the ***mediating effects*** of time-dependent covariates on group differences in the risk of an outcome. A mediating variable is one that also differs between groups and is a strong risk factor for the outcome. Ideally, a mediator (e.g., HbA1c) is a variable in the causal pathway between the exposure (e.g., treatment group) and the outcome (e.g., nephropathy), and is useful in explaining the mechanisms by which the exposure affects the outcome. Under Baron and Kenny's mediation paradigm (1986), three regression models are employed:

1. regressing the outcome on the exposure (e.g. treatment group);
2. regressing the mediator on the exposure; and
3. regressing the outcome on both the exposure and the mediator.

A change in the estimate of the exposure effect from model 1 to model 3 is evidence of mediation. More specifically, the total effect of the exposure on the outcome (Exposure → Outcome path) in model 1 can be decomposed in the direct effect (Exposure → Outcome path) in model 3 and the indirect effect (Exposure → Mediator → Outcome path) in model 3. Furthermore, the mediation proportion, defined as the proportion of the total effect explained by a particular mediator (i.e., the indirect effect divided by the total effect), will also be reported.

If the treatment group effect on the outcome is mediated (explained) by the treatment group effect on the time-dependent covariate, then the treatment group effect should become close to zero when adjusted for the time-dependent covariate, termed full mediation (Baron, Kenny, 1986). Partial mediation occurs when adjustment for the covariate results in a reduction of the group effect, but not its complete elimination. Further, it is possible that a covariate can be both an effect modifier and a mediator (MacKinnon, 2011). The unified decomposition of effects attributed to the mediation or interaction (VanderWeele, 2014), will provide insight into how much of the randomized intervention's effect can be attributed to mediation alone, interaction (effect modification) alone, or both/neither interaction and mediation.

With time-to-event data (e.g., time to death), the proportional hazards (PH) assumption is not preserved under marginalization (i.e., the PH assumption cannot hold for both models 1 and 3) (Gail et al, 1984). Instead, the Aalen additive hazards model will be used for the time-to-event outcome (i.e., models 1 and 3) (Martinussen, Scheike, 2010). When properly adjusted for confounders, the results of these mediation analyses also have causal interpretation (Lange, Hansen, 2001, Bebu et al, 2015, Vanderweele, 2015).

2.14 Subgroup and Stratified Analyses

Analyses will also be conducted assessing the differences in study outcomes within segments of the study cohort defined from characteristics assessed at baseline, such as by gender. For each stratification factor (e.g., gender), the treatment groups will be compared separately within each stratum or subgroup and then a test of homogeneity between strata (no stratum by group interaction) will be tested. Initially the within stratum and between strata tests will be conducted using a multivariate test of the equality (and homogeneity) of the differences among the 4 treatment groups simultaneously. If any heterogeneity is detected, then additional tests will be conducted separately for each of the 6 pair-wise drug group comparisons. Such tests can be conducted using an appropriate regression model for each outcome, such as a Cox proportional hazards model for the time to primary metabolic failure. For strata defined from a quantitative variable (e.g., age), an additional test of interaction will be conducted using the quantitative covariate rather than simply the discrete strata.

In this instance we have recently shown that the successive testing of homogeneity among subgroups and the subsequent testing within subgroups can be conducted under the closed testing principal. If there are only 2 strata, such as males and females, the test of homogeneity among the 4 groups across strata can be conducted at level α. If significant, then the test between the 4 groups can be tested at the same level α separately for males and separately for females, with no adjustment needed for the two tests. Suppose then that the 4-group test among males is significant at level α. Then we can proceed to test the sets of three group differences, and wherever 2 of these are significant at level α, the component pairwise test can also be conducted at level α.

The above also generalizes to the case where there are 3 or more strata. For the case of 3 strata, the homogeneity among all 3 strata is tested, and if significant at level α, the homogeneity of the next level sets for each pair of strata is tested at level α. If any two of these are significant then further testing of the group differences can be tested with the common stratum. For example, if the test of homogeneity of strata 1, 2, and 3 is significant then the homogeneity of 1 and 2, 1 and 3 and 2 and 2 are each tested. If the tests of 1 and 3 and 2 and 3 are significant, then within the third strata additional testing can be conducted comparing the treatment groups, starting with a test among the 4 groups, all testing at level α.

The baseline factors to be considered include race/ethnicity, gender, age, diabetes duration, weight, BMI, HbA1c, and measures of insulin sensitivity, insulin secretion, and the glucose disposal index, all measured at baseline or prior to randomization.

Age will be stratified as <45, 45-59, ≥60 years, and other quantitative covariates will be stratified by tertiles.

2.15 Risk Factor Modeling

Multivariate modeling will also be conducted to determine risk factors associated with study outcomes such as primary metabolic failure. These models will employ a large number of fixed baseline covariates and also time-dependent covariates over time, some representing either the current (most recent) measurement, or the updated mean of all follow-up values since randomization, or both. Along with input from clinicians, a comprehensive analysis of collinearity will be conducted (Belsley, 1991) to better understand potential confounding issues.

Given the large number of risk factors, variables will be entered into the models (e.g., a Cox proportional hazards model) one block at a time, starting with design factors, then demographic-physical, etc. The variable selection/deletion process will be guided by statistical significance (p-values), penalized likelihood (e.g., lasso) and the Akaike Information Criterion (AIC). The lasso achieves simultaneous estimation and variable selection by shrinking the regression coefficients toward zero, and setting those deemed unimportant to exactly zero (Tibshirani, 1997), while the AIC (Claeskens, Hjort, 2008) is a likelihood-based measure of model fit adjusted for the number of covariates in the model (lower is better). After adding each block, a variable will be deleted if not nominally significant, and yields a poor AIC value, and has a penalized likelihood estimate of zero. After the last block is entered, the final model is fit using the selected covariates. Two additional sensitivity analyses will start with the complete set of variables followed by subset selection based either on the AIC or the penalized likelihood (in a backwards selection fashion).

2.16 Composite Outcomes

In addition to the primary and secondary clinical outcomes based on failure to maintain metabolic control, composite outcomes will be assessed among treatment groups that reflect durability of glycemic control and tolerability to the assigned medications. Each composite outcome will consist of multiple component outcomes that will be assessed using multivariate analyses. For example, at 4 years of follow-up the differences among the treatment groups may be assessed in a multivariate composite outcome comprising the HbA1c level, weight, and any history of hypoglycemia to determine whether one treatment has a better outcome for all three components simultaneously than another treatment, i.e lower HbA1c, lower weight and freedom from hypoglycemia. This can be assessed using a multivariate one-sided (or one-directional) test, or a test of stochastic ordering as described in Section 2.9. One simple such test is that of O'Brien (Sparling et al, 2006) that is based on each subject's mean of the rank scores for each component. This is suitable for an analysis of multiple quantitative (or ordinal) components at a single point in time, but would not apply to the above composite that includes a binary outcome for hypoglycemia. Alternatively, the Wei-Lachin test of stochastic ordering (Wei, Lachin, 1984, Lachin, 1992, Lachin, 2014) could be used to conduct a one-directional test of the multiple components using a separate analysis for each component, such as a test of difference between means for one component, test for proportions or event-times for another, or a test of incidence rates for another (Lachin, 2014). For the composite above, the difference in the mean HbA1c, the mean weight at 4 years of follow-up, and the rate of hypoglycemia per 100 patient-years over the 4 years of follow-up, could be assessed jointly. An analysis with adjustment for baseline covariates can also be conducted (Lachin, 2014).

Some composite outcomes will consist of the times to multiple events, such as the time to CVD death, the time to non-fatal MI and the time to non-fatal stroke, where a given subject can experience all three outcomes successively. The traditional method for the analysis of such a composite is to employ the time to the first of any of the component events in a simple survival

analysis such as using a Cox PH model. In cardiovascular outcome trials this is termed the major adverse cardiovascular event (MACE) outcome. Recently, however, we have shown (Lachin, Bebu, 2015) that this simple approach can be less powerful that a Wei-Lachin one-directional test based on separate PH models for each component event, i.e. separate models for CV death, non-fatal MI and non-fatal stroke. The Wei-Lachin test is more powerful in part because it employs the times to all component events experienced by a subject, not just the first. The test is also based on the average log hazard ratio among the component events that provides a useful summary of the average treatment group difference over the set of three outcomes.

In addition, the incidence or prevalence of a composite outcome will be assessed using a single joint outcome, such as the proportion (prevalence) of subjects at 4 years who are still able to maintain an HbA1c <7% without having experienced any hypoglycemia or gained any weight, i.e. the proportion of subjects who satisfy all three criteria simultaneously. A longitudinal analysis will be conducted of the proportions meeting this criterion at each visit over time, and a "survival" analysis will also be conducted based on the time to failure to maintain this composite outcome (i.e., the time to either the primary outcome or hypoglycemia or weight gain).

3 PRIMARY METABOLIC OUTCOME ANALYSES

3.1 Data

Under the protocol, patients are scheduled to have quarterly clinic visits at which time a HbA1c is measured. From the successive HbA1c values, primary, secondary and tertiary metabolic outcomes are determined (described below).

It is preferable that these HbA1c measurements be conducted in the Central Biochemistry Laboratory (CBL). However, there are situations where a patient is unable to attend a clinic visit. In this case, the clinical center can provide a kit for the patient to use to draw a capillary collection that is then forwarded to the CBL for assay. The CBL HbA1c assay has been shown to provide almost identical values when conducted using a remote capillary collection which is then shipped to the CBL as with a venous sample obtained in the clinical center. In addition, there will be instances where a patient misses a visit with no contact (no capillary collection) but where a HbA1c may be available from the EMR system. In such instances the clinical center may also add the EMR HbA1c to the study data base.

In cases where the patient deviated from the quarterly visits and there is a question as to whether a metabolic outcome occurred, the history of HbA1c values, including capillary collections and possibly EMR values, may be considered by the Adjudication Committee to determine whether the outcome occurred. All such adjudicated outcomes would be included in the analyses of metabolic outcomes.

3.2 Intent-to-treat Analysis

The primary outcome is the time to observation of HbA1c $\geq$7% at a quarterly visit, with subsequent confirmation, while being treated at the maximum tolerable doses of both metformin (up to 2000 mg per day) and the randomly assigned medication. The initial HbA1c value $\geq$7% is termed the triggering value and the subsequent HbA1c value is termed the confirmation value. In order to ensure that all patients have adequate time for their regimen to reach the maximum tolerable dose, an HbA1c value $\geq$7% at the month 3 visit will not count towards the declaration of a primary metabolic outcome. That is, month 6 is the earliest that a triggering HbA1c value $\geq$7% can be observed. Since the declaration of the outcome requires two successive elevated HbA1c values, the assignment of a right censoring event time must account for the confirmation process as described previously in Section 2.4.3.

Using the intent-to-treat cohort, differences between groups will be tested, and relative risk estimates obtained from a Cox proportional hazards model for discrete time observations, adjusted for the baseline HbA1c with a four category class covariate to represent the four drug class groups (Lachin, 2011). See Section 2.4.3. A single overall omnibus Wald test at the 0.05 significance level will be conducted comparing the 4 drug combination groups. Significance tests and relative risk (hazard ratio) estimates for each of the four 3-group comparisons, and for the six pair-wise drug group comparisons will be obtained as contrasts among model coefficients from the overall 4 group Cox model. As described in Section 2.2 the closed testing principle will then be applied to determine statistical significance of each of the pair-wise comparisons. For the comparison of groups 1 vs 2 to be declared significant at level 0.05, the 4-group test must be significant at that level, as well as the 1,2,3 and 1,2,4 three group comparisons and then the 1 versus 2 comparison.

Some patients will deviate from the protocol schedule of HbA1c assessments either through missed visits following a triggering HbA1c, or failure to use the central laboratory for HbA1c measures. In such cases the site should attempt to obtain other HbA1c measures when available such as through an EMR system. The Adjudication Committee will review the available data from each such subject to determine whether a primary metabolic outcome occurred and if so when, either as a point in time or as an interval of time. In the latter case the midpoint of the interval will be employed as the event time.

If there are a substantial number of such patients (> 100 total) where an interval of time is provided by the Adjudication Committee, then a Weibull regression model for interval censored data (Sparling et al. 2006) will be employed in lieu of the Cox PH model.

Other patients may deviate from the protocol schedule of diabetes medications and start other medications than that originally assigned at randomization prior to the occurrence of the confirmed primary metabolic outcome. These events would be counted in the intent-to-treat analysis. However, a sensitivity analysis will be conducted in which the subject's time to the primary outcome is right censored at the time that the deviation occurs in the above Cox or Weibull models. A sensitivity analysis will also then be conducted using these time to protocol deviations as a competing risk and the groups compared using the Fine and Gray (1999) model for the cumulative incidence function of the time to primary outcome in the presence of deviation as a competing risk (see Section 2.4.5).

Finally, if there is a difference among groups in mortality, then further sensitivity analyses will be conducted using mortality as a competing risk in the evaluation of the cumulative incidence functions among groups. See Section 2.4.5.

3.3 Secondary Analyses of the Primary Outcome

3.3.1 Modification and Mediation

If the primary analyses indicate that one or more treatments are superior to other treatments then additional analyses will be conducted to assess whether any of the fixed baseline covariates can be considered an effect modifier. This would be assessed by a test of a treatment group by covariate interaction in the Cox PH model of the primary outcome.

Additional analyses will also attempt to identify the factor(s) that may mediate the beneficial effect of a given treatment. This will be assessed using time-dependent covariates that reflect the effects of treatment on other factors over time, principally changes in the oral disposition index using measures of insulin resistance and insulin secretion obtained from the oral glucose tolerance test (OGTT) performed at baseline and again at 1, 3 and 5 years.

These analyses would be conducted as described in Section 2.14.

3.3.2 Subgroup Analyses

Further analyses will examine treatment group differences in the primary outcome among subgroups defined as subsets of the total cohort that are stratified by specific baseline factors. The testing of treatment group differences between and within subgroups will be conducted using the closed testing principle as described in Section 2.15 that also specifies the minimum set of stratification factors to be employed.

3.3.3 Risk Factors

The multitude of factors measured at baseline and during follow-up, including physiologic and mechanistic variables, such as those from the OGTT, will be employed in additional analyses to identify factors that are associated with an increased risk of primary metabolic failure. If one or more treatments are declared superior to the others, then separate risk factor models will be developed within each superior treatment and the others combined. See Section 2.16.

Exploring these variables may provide information to help clinicians in selecting the medication that will work best for that individual patient. This will promote a more detailed understanding of the mechanisms by which the drug classes do or do not prolong the time to such glycemic deteriorations, and to define different metabolic phenotypes with varying risk of such deterioration.

4 SECONDARY AND TERTIARY METABOLIC OUTCOMES AND ANALYSES

4.1 Secondary and Tertiary Metabolic Failure

The time to secondary metabolic failure (defined as HbA1c >7.5% with subsequent confirmation) after achieving the primary outcome and while receiving the maximally tolerated dose of the assigned regimen. The primary and secondary outcomes may be reached simultaneously if the initial value and the confirmation are both >7.5%.

Among those randomly assigned to the three non-insulin groups (glimepiride, sitagliptin and liraglutide), after the secondary outcome has been confirmed the subject then begins to receive the addition of insulin glargine to metformin and the randomly assigned therapy. After being on glargine for a minimum of 60 days, the subject is then at risk of tertiary metabolic failure defined as an HbA1c >7.5% subsequently confirmed. If the tertiary outcome is reached and confirmed, the randomly assigned treatment is withdrawn and the subject is then administered an intensive insulin regimen using both glargine and rapid acting insulin.

Among those randomly assigned to glargine, after the secondary outcome has been reached the subject continues metformin and glargine and is administered the intensive insulin regimen with rapid-acting insulin.

Thus, additional time-to-event outcomes can be defined:

a. Time from randomization to the secondary metabolic failure (triggering HbA1c subsequently confirmed).
b. Time from primary failure confirmation to the secondary metabolic failure trigger. Those who trigger or confirm the primary outcome at the same time as the secondary outcome would have the event time zero.
c. Time from randomization to the tertiary metabolic failure (triggering HbA1c subsequently confirmed) among those originally assigned to glimepiride, sitagliptin or liraglutide.
d. Time from the secondary failure confirmation to the tertiary metabolic failure (triggering HbA1c subsequently confirmed) among those originally assigned to glimepiride, sitagliptin or liraglutide.

e. Time from randomization to the initiation of intensive insulin therapy in all four treatment groups.

The analyses described in Section 3.2 for the primary outcome will also be applied to each of these outcomes.

4.2 Other Metabolic Outcomes and Analyses

a. The proportion over time of participants among treatment groups who experience the primary metabolic failure (HbA1c $\geq$7%).

   At each quarterly visit the hazard rate of incident primary metabolic failures (proportion of failures among those at risk) is computed and a smoothed hazard rate function estimated over the complete follow-up period. At a given visit the failures is the number who triggered for the primary outcome at that visit that was subsequently confirmed, and the number at risk is the number followed at that visit less those who have reached the primary outcome at a previous visit (triggered and subsequently confirmed).

b. The proportion over time of participants among treatment groups who experience the secondary metabolic failure (HbA1c >7.5%).

   Same analyses as 4.2.a above.

c. The proportion over time of participants among the non-insulin treatment groups who experience the tertiary metabolic failure (HbA1c >7.5%).

   Same analyses as 4.2.a above.

d. The cumulative incidence of the implementation of glargine therapy (defined as basal insulin), while being treated at maximum tolerable doses of the assigned regimen.

   Same analyses as in Section 3.1 This event will include cases that start glargine insulin therapy off protocol and regardless of whether or not a secondary outcome has been declared.

e. The proportion over time of participants among treatment groups who had glargine insulin therapy initiated.

   This event will include cases that start glargine insulin therapy off protocol and regardless of whether or not a secondary outcome has been declared. See Section 4.2.a.

f. The cumulative incidence of the implementation of intensive insulin therapy (defined as basal plus rapid-acting insulin), while being treated at maximum tolerable doses of the assigned regimen.

   Same analyses as in Section 3.2 This event will include cases that start intensive insulin therapy off protocol and regardless of whether or not a tertiary outcome has been declared.

g. The proportion over time of participants among treatment groups who had intensive insulin therapy initiated.

   See Section 4.2.a.

h. Proportional rate model of the differences between groups in the incidence of primary, then secondary and then tertiary failure with a plot of the successive cumulative incidence functions with time 0 the day of randomization.

i. Stratified recurrence models of the differences between groups in the time from randomization to primary (time 0 = randomization), from primary to secondary (time 0 = primary trigger) and from secondary to tertiary (time 0 = secondary confirmation), with a plot of the three cumulative incidences. See Section 2.5.

j. Proportional rate model of the differences between groups in the incidence of primary outcome, then glargine, then intensive insulin therapy with a plot of the successive cumulative incidence functions with time 0 the day of randomization.

k. Stratified recurrence models of the differences between groups in the time from randomization to primary (time 0 = randomization), from primary to glargine (time 0 = primary trigger) and from glargine to intensive therapy (time 0 = date start glargine), with a plot of the three cumulative incidences.

5 OTHER DIABETES-RELATED OUTCOMES AND ANALYSES

5.1 Data

HbA1c data are described above. Fasting plasma glucose (FPG) is measured annually at the CBL.

The OGTT is measured at baseline and years 1, 3 and 5. This includes a basal (fasting) collection (time 0) and collections periodically up to 2 hours after drinking a liquid meal during which glucose, insulin and C-peptide are measured. From these measurements, indices of insulin resistance and insulin secretion are computed, principally the inverse insulin and insulinogenic index, respectively, the latter being the time 0 to 30 minute change in insulin divided by the like change in glucose, also called the insulin to glucose ratio (IGR).

Severe hypoglycemia is defined as an episode in which the patient was unable to treat the event on his/her own and required the assistance of another individual. There are three additional classifications of severe hypoglycemia based on whether or not:

- the episode was accompanied by coma and/or seizure (major hypoglycemia);
- the episode led to injury of the patient or others;
- the episode was accompanied by a motor vehicle accident in which the patient was the driver.

The date of all such episodes is recorded.

In addition, at every quarterly visit the patient indicates whether any (one or more) episode of *symptoms of* hypoglycemia occurred within the past 30 days prior to the visit, as well as characteristics of these episode(s).

Cognitive function will be assessed at baseline and years 4 and 6 of follow-up using a battery of neurocognitive tests that are scored centrally.

5.2 Quantitative Measurements

The following analyses will be conducted for HbA1c, FPG, and measurements derived from the OGTT to assess insulin resistance and beta-cell function.

a. For each measurement, a longitudinal repeated measures analysis of the mean values over the study duration along with an estimate of the average value over all visits (and the area under the curve), and the assessment of treatment group differences, adjusted for the baseline value (See Section 2.9).
b. For each measurement, estimate the mean change in HbA1c from baseline to year 4 in an ANCOVA model adjusted for the baseline value (2.8).

5.3 Hypoglycemia

a. Crude rates per 100 patient years of confirmed symptomatic hypoglycemia (defined as relieved by food and/or with BG<70 mg/dl) will be computed for each year of follow-up and over all years combined. Treatment group differences will be assessed using an appropriate model for count data (Poisson, Negative binomial, etc.). (See Section 2.6).
b. Analyses as in 5.4.a will also be applied to rates of

- severe hypoglycemia defined as requiring third party assistance;
- severe hypoglycemia episodes resulting in coma and/or seizure.
- Severe episodes resulting in injury of the patient or others
- Severe episodes accompanied by a motor vehicle accident in which the patient was the driver
- severe hypoglycemia satisfying any of these three criteria.

c. Analyses of recurrent severe hypoglycemia will be conducted using the methods of Section 2.5 to assess group differences within subgroups as described in Section 3.3.3, analyses of risk factors described in Section 3.3.4 and analyses of mediation factors as in Section 3.3.2.

d. Analyses of recurrent major hypoglycemia will also be conducted as in Section 5.4.c.

5.4 Cognition

a. Each element of the neurocognitive battery will be analyzed longitudinally using the methods described for the analysis of ordinal data in Section 2.7, and an aggregate or composite analysis conducted as described for the analysis of composite outcomes in Section 2.16.

6 CARDIOVASCULAR EVENTS AND RISK FACTOR ASSESSMENTS

6.1 Data

Cardiovascular risk factors are routinely measured during follow-up, lipids annually, physical assessments (blood pressure and weight) quarterly, and medication use (antihypertensive, cardio-protective) quarterly.

Body weight, waist circumference, hip circumference, and body mass index (BMI) are measured quarterly. Obesity will be defined as BMI$\geq$30 kg/m^2, and major obesity as BMI$\geq$35 kg/m^2.

Hypertension is defined as blood pressure $\geq$140 mmHg systolic, $\geq$90 mmHg diastolic, or use of blood pressure lowering medications for control of blood pressure. Hyperlipidemia is defined as LDL cholesterol levels $\geq$100 mg/dl or the use of lipid-lowering medications.

Cardiovascular outcome events are recorded in real time and then adjudicated by the study Adjudication Committee that includes experts from outside the study. The primary CVD outcomes are the components of MACE – cardiovascular death, nonfatal MI, nonfatal stroke. Other events of interest are silent MI on ECG, unstable angina and revascularization. Congestive heart failure requiring hospitalization will also be recorded.

An ECG will be conducted at baseline and years 2, 4 and 6 of follow-up and will be read centrally to report the presence or absence of any, minor and/or major abnormalities. Specific ECG-detected abnormalities will also be reported including silent MI, myocardial ischemia, left ventricular hypertrophy, arrhythmias, and conduction defect. Presence of cardiac autonomic dysfunction will also be reported.

Participants satisfying the criteria for obesity, hypertension, hyperlipidemia, abnormal ECG and CAN at baseline will be excluded from the analyses of incidence of each respective outcome.

6.2 Quantitative Measurements

The analyses described in Section 5.2 above will also be applied to total cholesterol, triglycerides, LDL, HDL and non-HDL cholesterol; systolic and diastolic blood pressure and

pulse pressure; and body weight, waist circumference, hip circumference, and body mass index (BMI).

These analyses will also be applied to longitudinal analyses of the estimated CVD risk calculated using the UKPDS, Framingham or other cardiovascular risk engine (D'Agostino et al, 2008, Stevens et al, 2001).

6.3 Obesity

For the analysis of obesity or major obesity, participants with the condition at baseline will be excluded from analysis.

a. The cumulative incidence of the time to the development of obesity (BMI$\geq$30 kg/m^2) and related analyses will be applied as described in Section 3.2.
b. The proportion of subjects who develop obesity at each annual visit (hazard rate) will be described using the analyses as in Section 4.2.a.
c. The prevalence of obesity at each annual visit will be computed and a longitudinal repeated measures (GEE) analysis of the prevalence over the study duration performed, along with an estimate of the average prevalence over all visits, and the assessment of treatment group differences, adjusted for the baseline value (See Section 2.3).
d. Analyses of the incidence of obesity will also be conducted within subgroups as described in Section 3.3.3 as will analyses of risk factors described in Section 3.3.4 and analyses of mediation factors as in Section 3.3.2.
e. Analyses as in 5.3.a and 5.3.b will also be applied for the assessment of major obesity (BMI$\geq$35 kg/m^2).

6.4 Other Binary Outcomes

The analyses described above in 6.3.a, b and c will also be conducted for:

a. The use of blood pressure lowering agents over time.
b. Hypertension.
c. Emergent hypertension among those who had levels <140/90 and were free of blood pressure-lowering medication use at baseline.
d. Use of drugs to treat dyslipidemia.
e. Hyperlipidemia. Participants with hyperlipidemia at baseline will be excluded.
f. Any ECG abnormality (minor or major), those with abnormality present at baseline excluded.
g. Any major ECG abnormality, those with a major abnormality present at baseline excluded.

6.5 Cardiovascular Events.

a. Incidence of CV death, non-fatal MI, and non-fatal, each conducted separately (See Section 2.4.1) along with a joint Wei-Lachin analysis (Section 2.16)
b. Incidence and prevalence of ECG-detected abnormalities, including silent MI, myocardial ischemia, left ventricular hypertrophy, arrhythmias, and conduction defect (Sections 2.4.2 and 2.3).
c. Incidence and prevalence of cardiac autonomic dysfunction (Sections 2.4.2 and 2.3).
d. Incidence and prevalence of other cardiovascular events including unstable angina requiring hospitalization or revascularization (Sections 2.4.1 and 2.3).
e. Incidence and prevalence of congestive heart failure requiring hospitalization (See Sections 2.4.1 and 2.3).

7 SECONDARY MICROVASCULAR OUTCOMES

7.1 Data

The urinary albumin:creatinine ratio (ACR) will be measured 6-monthly. The presence of microalbuminuria (or worse) will be defined as a value >30 mg/g, and the presence of albuminuria defined as a value >300 mg/g.

Serum creatinine is measured annually from which an estimated GFR (eGFR) will be computed using the EPI-CKD formula (Inker, 2012). Impaired GFR will be defined as an eGFR < 60 mL/min per 1.73m^2.

Peripheral neuropathy will be classified as present or absent based on the Michigan Neuropathy Screening Instrument administered annually.

Cardiac Autonomic Neuropathy will be classified as present or absent based on the central grading of the ECG conducted at baseline and years 2, 4 and 6 of follow-up.

A history of retinal photocoagulation and other ocular procedures will be collected quarterly.

Participants with an outcome of interest at baseline will be excluded from the analyses of the incidence of that event during the study.

7.2 Analyses

a. The analyses in Section 5.2 will be applied to the longitudinal analysis of the albumin:creatinine ratio values over time.
b. The incidence and prevalence of microalbuminuria among subjects who had ACR levels <30 mg/g at baseline will be assessed as in Sections 2.4.2 and 2.3.
c. The incidence and prevalence of macroalbuminuria among subjects who had ACR levels <300 mg/g at baseline will be assessed as in Sections 2.4.2 and 2.3.
d. The analyses in Section 5.2 will be applied to the longitudinal analysis of the eGFR values over time. Analyses of eGFR using random effects models as in Section 2.10 will also be applied to the rate of decline of eGFR over time.
e. The incidence of renal insufficiency will be assessed as in Section 2.4.2.
f. The incidence of retinal photocoagulation for diabetic retinopathy and other ophthalmologic procedures by quarterly self-report will be assessed as in Sections 2.4.2 and 2.3.
g. The incidence of peripheral neuropathy will be assessed as in Sections 2.4.2 and 2.3.
h. The incidence of CAN on ECG will be assessed as in Sections 2.4.2 and 2.3.

8 ADVERSE EFFECTS

Serious adverse events and adverse events of special interest are recorded as they occur. The following analyses will be performed.

a. The incidence of pancreatitis, pancreatic and medullary thyroid cancer, other cancer types (except non-melanoma skin cancer) will be summarized as the number of subjects affected, the number of events (episodes), and the rate of events using the methods described in Sections 2.3 and 2.6. If adequate numbers of such affected subjects (at minimum 50 in total), additional analyses as in Sections 2.4.1 and 2.5 may also be conducted.
b. Like analyses will be conducted for the incidence of severe adverse events using the higher level MedDRA terms/codes of the event types.
c. Like analyses will be conducted of the incidence of hospital admission.

9 ADHERENCE-TOLERABILITY

Adherence to, and tolerance of, the study medications over the study duration will be recorded. These are principally reflected by the incidence of changes in dose or withdrawal of medications. Treatment satisfaction will be periodically assessed by subject ratings using a standardized questionnaire (DTSQ). Other indices of adherence include taking non-study glucose-lowering medications and failure to take assigned study medications. The following analyses will be performed.

a. The incidence of intolerance to medications indicated by modification of the dose of metformin or other agents in response to gastrointestinal symptoms will be assessed as in Section 8.a.
b. The incidence of intolerance to medications indicated by modification of the dose of metformin or other agents in response to other symptoms will be assessed as in Section 8.a.
c. The incidence of intolerance to medications indicated by modification of the dose of metformin or other agents specifically in response to hypoglycemia will be assessed as in Section 8.a.
d. Treatment satisfaction over study duration will be examined using methods for the analysis of ordinal data as described in Section 2.7.
e. Other metrics of treatment satisfaction include the numbers of subjects taking other (non-study) glucose-lowering medications and the numbers who failed to take the assigned medications, and the period of time of each.

10 HEALTH ECONOMIC ANALYSES

Costs related to treatment of diabetes and its morbidities will be recorded. Quality of life will also be assessed that will allow the computation of utilities and the data summarized as quality adjusted life years and other metrics such as the cost-effectiveness ratio.

a. Quality of life over study duration will be examined using methods for the analysis of longitudinal ordinal data as described in Section 2.7.
b. Incidence of all-cause mortality will be described using the survival analysis methods of Section 2.4.1.
c. Incidence of diabetes related mortality, and non-diabetes related mortality, will likewise by assessed.

11 SECONDARY COMPOSITE OUTCOMES

A secondary composite outcome refers to analyses intended to assess whether one treatment produces greater benefit for a collection of outcomes. The analysis starts with an analysis of each component outcome separately. The results of those analyses will then be used with the Wei-Lachin multivariate one-directional analysis to test differences between groups for a preponderance of benefit over the set of outcomes in that composite as described in Section 2.16. This approach will be applied to the following composite outcomes.

a. The mean HbA1c, the mean body weight, and the rate per year of severe hypoglycemia since randomization over time using an appropriate longitudinal model for each component outcome.
b. The mean HbA1c, the mean body weight, and the rate per year of severe hypoglycemia since randomization at 4 years of follow-up.

c. An analysis as in 11.a will also be conducted for HbA1c and weight alone.
d. An analysis as in 11.a will be conducted for the time to an episode of severe hypoglycemia and the time to the primary metabolic outcome.
e. An analysis as in 11.a will be conducted for the time to an episode of severe hypoglycemia and the time to weight gain of 5% or more of baseline body weight.
f. A longitudinal analysis of the proportion of participants that both have not yet reached the primary metabolic outcome without having experienced any hypoglycemia over time and without any weight gain over time using the methods described in Section 2.3.

12 SAMPLE SIZE AND STUDY POWER

12.1 The Primary Metabolic Outcome.

The incidence of reaching the primary metabolic outcome will be compared among groups using a Mantel-logrank test under a proportional hazards model. The original protocol specification at the start of GRADE was to enroll 5000 patients over 3 years assuming a lagged (concave) recruitment pattern in which 40% are enrolled over the first 18 months and 60% enrolled over the final 18 months using the method of Lachin and Foulkes (1986). A total trial duration of 7 years would then provide 90% power to detect a 25% difference (hazard ratio of 0.75) between any two of the four study agents in the risk of the primary metabolic outcome assuming a hazard rate of 0.0875 / year in the less effective group, and losses to follow-up at 4% per year, adjusted for 6 pair-wise tests.

A key assumption in this assessment is the projection of a hazard rate of 0.0875 per year for the primary outcome that was initially based on a conservative estimate obtained from the ADOPT study (Kahn et al., 2006) that compared metabolic control among subjects assigned to sulfonylurea, metformin or rosiglitazone in a population comparable to that in GRADE. If the actual hazard rate is higher or lower than this estimate, then the power of the study to detect a 25% risk reduction will be greater or less than the 90% computed above. The GRADE Data and Safety Monitoring Board (DSMB) appointed by the NIDDK is charged with monitoring not only the safety of the trial but also its overall feasibility. The DSMB recommended to the NIDDK that the actual observed hazard rate in the study be masked to the study investigators, a recommendation with which the NIDDK concurred. One of the many reasons is that we would still like to address the long-term microvascular and macrovascular outcomes for which the full sample size is desired. Thus, the investigators will not reassess the design assumptions that led to the calculation supporting the need for 5000 patients.

Recruitment, however, has lagged. In May, 2016 the NIDDK and the DSMB requested that we conduct a detailed examination of the power of the study using the above initial assumptions but allowing for the actual pattern of recruitment observed to date, and allowing for a 6 month extension of the recruitment period to 3.5 years, and a like extension of the total study duration from 7 to to 7.5 years, with the expectation that the study might still fall short of the goal of 5000 patients. Over the first 1.4 years of recruitment the study enrolled 1550 patients at a rate of about 92 per month. Thereafter the pattern of recruitment has been strongly linear with a rate of about 130.4 per month. At this rate projected to the end of the 3.5 year recruitment period, 4836 subjects would be randomized .

Because the recruitment was strongly linear at 92 subjects/month over the first 1.4 years, and 130/month thereafter, rather than employ a concave recruitment pattern these updated power computations used a stratified calculation with 2 strata, one that enrolled 92 / month over 1.4 years who were followed for up to 7.5 years, and another that enrolled 130.4 / month over 2.1 years and followed for up to 6.1 years. The total of 4836 would yield 90.5% power, virtually identical to the original protocol computation owing to both the 6 month extension and the linear

rate of recruitment. On this basis the NIDDK, with the concurrence of the DSMB, approved the extension of the study recruitment and follow-up periods to 3.5 and 7.5 years, respectively.

However, recruitment recently slacked for a couple months and it is possible that the final enrollment would be only 4600 subjects. In this case, the power would be reduced slightly to 88.9%.

Recruitment will close by May 1, 2017 at which time the final study power will be determined for all outcomes under the original protocol design assumptions along with the 3.5/7.5 recruitment and follow-up periods and the final achieved sample size.

The following sections present the original power computations from the initial study protocol for the other secondary outcomes and for analyses among subgroups. Those calculations assumed 5000 patients enrolled over 3 years and followed up to 7 years of major interest. As described above, the final levels of power with a final sample size of 4600 to 4800 will be close to these calculations from the original protocol.

12.2 Secondary Outcomes – Microalbuminuria and Clinical Cardiovascular Disease

The incidence of onset of microalbuminuria on a biannual measure of the urinary albumin/creatinine ratio will be compared among groups using a Mantel-logrank test under a proportional hazards model. From other studies, the hazard rate of onset of microalbuminuria is projected to be about 0.04 per year in whichever group has a higher event rate (Lachin et al., 2011). For the 4-way comparison among the 4800-5000 subjects, the study would have 88% power with a hazard rate of 0.04/year, 92% with 0.045/year, to detect a 33% difference in risk for microalbuminuria between any pair of groups.

In the ADOPT study (Kahn et al., 2006), the incidence of MACE was 0.76% per year and of MACE plus congestive heart failure was 1.14% per year. Assuming a more conservative incidence rate of 1% per year and the other assumptions above, GRADE will provide 80% power to detect a 50% difference in the risk of CVD between any pair of drug groups, adjusted for 6 pair-wise comparisons. The study also has 80% power to detect a 42% difference in risk in an analysis of each drug group compared to all other drug groups combined, adjusted for 4 comparisons.

12.3 Subgroup Analyses

Assume that in the overall study one drug group has a hazard ratio for the primary outcome of 0.75 versus the other three drug groups. Using the methods of Lachin (2013), for a test of homogeneity of the 4-way drug group difference within two equal sized strata (subgroups) of 2500 subjects each, the study will provide 94% power to detect a pattern of drug group differences where the hazard ratio is 25% greater (HR = 0.938) within one stratum and 25% less (0.563) in the other. For the case of three strata with 1667 subjects each, the study provides 69% power to detect heterogeneity of hazard ratios of 0.563, 0.75, and 0.938.

13 ADDITIONAL STATISTICAL ACTIVITIES

13.1 Publication Generation and Policy

All members of the GRADE Research Group and collaborators (e.g. ancillary study investigators) must follow the GRADE Policy on Publications and Presentations. This is widely available through the GRADE website (https://grade.bsc.gwu.edu/web/GRADE/), under the tab for “Potential Collaborators”. The goals of this publication policy are to ensure the production of high quality manuscripts representing the scientific output of the Study; preserve the scientific integrity of the study in publications and presentations; protect the rights and privacy of the subject participants; provide the opportunity for members of the Research Group to participate

as authors of publications and presentations; and 3) provide appropriate and equitable authorship to those involved in the development, analysis, and writing of manuscripts.

Any member of the research group can propose a publication topic that is then reviewed by the Publication and Presentations Subcommittee (PPC). If approved, members of the Research Group are invited to volunteer to join the writing committee for the paper. The PPC meets periodically to evaluate the progress of analysis and writing of all papers. When the manuscript is complete it is then distributed to the members of the PPC for review. When approved by the PPC the paper is then distributed to the Research Group for review and approval.

GRADE has defined four categories of publications, each with specific authorship guidelines. Briefly, Category 1 manuscripts report the primary findings of the study with group authorship by the GRADE Research Group. Category 2 manuscripts focus on lesser findings in which the authors are named and includes “and the GRADE Research Group” with a citation to the list of group members. Category 3 papers report the results of ancillary studies with authorship as for category 2 except that the members of the ancillary study research group may also be cited as “and the “NAME” Study Research Group”. Finally, Category 4 manuscripts represent methodological papers and are authored by the writing team with acknowledgement of the support of GRADE. Responsibility for the category assignment for all manuscripts rests with the Publications and Presentations Committee, in consultation with the Executive Committee.

13.2 Current Publication Activity

The following is the list of publications generated by the GRADE research group to date.

1. Lachin JM. Sample size and power for a logrank test and Cox proportional hazards model with multiple groups and strata, or a quantitative covariate with multiple strata. Stat Med. 2013 Nov 10;32(25):4413-25. doi: 10.1002/sim.5839. PubMed PMID: 23670965; PubMed Central PMCID: PMC3775959.
2. Nathan DM, Buse JB, Kahn SE, Krause-Steinrauf H, Larkin ME, Staten M, Wexler D, Lachin JM; GRADE Study Research Group.. Rationale and design of the glycemia reduction approaches in diabetes: a comparative effectiveness study (GRADE). Diabetes Care. 2013 Aug;36(8):2254-61. doi: 10.2337/dc13-0356. PubMed PMID: 23690531; PubMed Central PMCID: PMC3714493.
3. Lachin JM. Applications of the Wei-Lachin multivariate one-sided test for multiple outcomes on possibly different scales. PLoS One. 2014 Oct 17;9(10):e108784. doi: 10.1371/journal.pone.0108784. PubMed PMID: 25329662; PubMed Central PMCID: PMC4201485.
4. Lachin JM, Bebu I. Application of the Wei-Lachin multivariate one- directional test to multiple event-time outcomes. Clin Trials. 2015 Dec;12(6):627-33. doi: 10.1177/1740774515601027. PubMed PMID: 26336199; PubMed Central PMCID: PMC4562325.
5. Lachin JM. Fallacies of last observation carried forward analyses. Clin Trials. 2016 Apr;13(2):161-8. doi: 10.1177/1740774515602688. PubMed PMID: 26400875; PubMed Central PMCID: PMC4785044.

Paper 2 presents the design of the study. The other 4 papers describe the development of statistical methods for application to GRADE. Paper 1 describes the multiple group methods to evaluate the sample size for the study. Papers 3 and 4 describe the application of the Wei-Lachin multivariate one-directional analysis to multiple (composite) outcomes and to multiple event time analyses for application in GRADE, and paper 5 describes the fallacies of a commonly used method to justify its exclusion from the GRADE study analysis plan.

In addition, the GRADE research group is looking forward to closure of the baseline data base following the close of recruitment as of May 1, 2017. Seven writing groups have been appointed, each of which may prepare multiple manuscripts based on the baseline data in each of the following domains: description of the study population, prevalent complications, recruitment practices, metabolic assessments, metformin, neurocognition and cost-effectiveness. These writing groups are now developing plans for specific analyses to be conducted when the data are closed.

The research group is masked to all outcome data, but in anticipation of the final analyses in 2021-22 additional writing groups will be appointed to start planning for end of study papers well before the data are closed.

13.3 Support of Ancillary Study Collaborations

The GRADE public website, https://grade.bsc.gwu.edu/web/GRADE/, includes links to the GRADE Publications and Presentations Policy, the Ancillary Studies and Sub-studies Policy and a description of the Ancillary Study or Sub-study Application Process. GRADE welcomes proposals from collaborators within or outside of the Research Group for research to address objectives and employ methodologies outside the scope of the core GRADE protocol. All applications are reviewed by the GRADE Ancillary and Sub-study Committee (ASC) in a timely fashion, and if approved, by the Research Group and the Data and Safety Monitoring Board.

Ancillary and Sub-studies differ in the extent of support and involvement of the GRADE infrastructure, especially the Biostatistical Research Center (BRC). An Ancillary Study is organized and managed by the ancillary study investigators, including data collection and analysis by the ancillary study research team. The BRC provides some oversight and may be required to provide access to some phenotypic data (after completion of the study) but otherwise will not be involved. A Sub-study is organized and managed by the GRADE BRC and GRADE investigators in collaboration with the sub-study research team and jointly seek independent funding.

Both ancillary and sub-studies require separate independent funding. GRADE investigators at all levels may collaborate closely with the ancillary and sub-study investigators to help secure independent funding to support the additional research. This funding includes all the procedural preparations to launch the study. In addition, the senior BRC statisticians assist in the design of the study and the specifications of the statistical considerations including the determination of sample size and power. The application is then submitted to the funding authority (usually NIH) with a budget that includes funds for the BRC to provide additional support such as data management, statistical analysis, and the construction of data files with phenotypic data required for analyses of the ancillary study data. In addition, the ancillary or sub-study funding will need to include support for the study coordinators, and possibly the clinical site PIs, if they will be required to participate in performing procedures or collecting data specifically related to the study.

In some cases, the ancillary study investigator may engage local statistical resources to conduct the data analyses in which case the BRC statisticians will provide oversight of the statistical activities conducted by the collaborator's statistician. The BRC statisticians, however, will provide statistical support for sub-studies.

13.4 Data Reports

The BRC generates reports describing the current state of the study for the biannual meetings of the Research Group and coordinators, and for the operational study committees. In addition, the BRC generates a report on the current state of the cohort and all study outcomes including tolerability and adverse events for review by the Data and Safety Monitoring Board.

empanelled by the NIDDK.

14 REFERENCES

Agresti, A. *Categorical Data Analysis*. New York: John Wiley and Sons, 1990.

Al-Khalidi HR, Hong Y, Fleming TR, Therneau TM. Insights on the robust variance estimator under recurrent-events model. *Biometrics* 2011; 67:1564-72.

Andersen PK, Borgan O, Gill RD, and Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag: New York, 1993.

Andersen PK, Borgan O, Gill RD, and Keiding N. Linear nonparametric tests for comparison of counting processes, with applications to censored survival data, (with discussion). *Int Stat Review* 1982; 50:219-44.

Andersen PK and Gill RD. Cox's regression model for counting processes: A large sample study. *Ann Stat* 1982; 10:1100-20.

Baron, RM, Kenny, DA. The moderator-mediator variable distinction in social psychological research. *J Pers Soc Psychol* 1986; 51:1173–82.

Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986; 51:1173-82.

Bebu I, Lachin JM. Large sample inference for a composite analysis based on prioritized components. *Biostatistics* 2016; 17:178-87.

Bebu I, Mathew T, Agan B. Inference for Surrogate Endpoint Validation in the Binary Case. *J Biopharm Stat* 2015; 25:1272-84.

Belsley DA. *Conditioning diagnostics: collinearity and weak data in regression.* New York: J. Wiley, 1991.

Bennett S. Analysis of survival data by the proportional odds model. *Stat Med* 1983; 2:273-77.

Bernardo MV, Harrington DP. Sample size calculations for the two-sample problem using the multiplicative intensity model. *Stat Med* 2001; 20:557-79.

Beyersmann J, Allignol A, Schumacher M. *Competing risks and multistate models with R*, Springer, 2012.

Boruvka, A. and Cook, R. J. A Cox-Aalen model for interval-censored data. *Scan J Stats* 2015; 42:414–26.

Cai J, Zend, S. Sample size/power calculation for case-cohort studies. *Biometrics* 2004; 60:1015-24.

Cantor AB. Sample-size calculations for Cohen's kappa. *Psychol Methods* 1996; 1:150-53.

Carpenter JR, Kenward MG. Multiple imputation and its application, Wiley, 2013.

Chen BE, Cook RJ, Lawless JF, Zhan M. Statistical methods for multivariate interval-censored recurrent events. *Stat Med* 2005; 24:671-91.

Chi GYH. Multiple testings: Multiple comparisons and multiple endpoints. *Drug Inf J* 1998; 32:1347S-1362S.

Claeskens G, Hjort NL. Model selection and model averaging. Cambridge; New York: Cambridge University Press, 2008.

Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*, 1972; 34:187-220.

Cox DR, Miller, HD. *The Theory of Stochastic Processes*. London: Chapman and Hall, 1965.

D'Agostino RB, Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care. *Circulation.* 2008;**117:**743-753.

DCCT Research Group. Comparison of Study Populations in the Diabetes Control and Complications Trial and the Wisconsin Epidemiologic Study of Diabetic Retinopathy. *Arch Int Med* 1995; 155:745-54. PMID:7695463.
DCCT Research Group. Hypoglycemia in the Diabetes Control and Complications Trial. *Diabetes* 1997; 45:271-86. PMID:9000705.
DCCT/EDIC Research Group. Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy. *N Engl J Med* 2000; 342:381-89. PMC2630213.
DCCT/EDIC Research Group. Effect of intensive therapy on the microvascular complications of type 1 diabetes mellitus. *JAMA* 2002; 287:2563-69. PMC2622728.
DCCT/EDIC Research Group. Intensive diabetes treatment and cardiovascular disease in type 1 diabetes mellitus. *N Engl J Med* 2005; 353:2643-53. PMC2637991.
DCCT/EDIC Research Group (Writing Committee: de Boer IH, Sun W, Gao X, Cleary PA, Lachin JM, Molitch M, Steffes MW, Zinman B). Effect of intensive diabetes treatment on albuminuria in type 1 diabetes: long-term follow-up of the Diabetes Control and Complications Trial and Epidemiology of Diabetes Interventions and Complications study. *Lancet Diabetes Endocrinol* 2014; 2:793-800. PMC4215637.
DCCT/EDIC Research Group (Writing Committee: Orchard TJ, Nathan DM, Zinman B, Cleary P, Brillon D, Backlund JC, Lachin JM). Association between 7 years of intensive treatment of type 1 diabetes and long-term mortality. *JAMA* 2015; 313:45-53. PMC4306335.
DCCT/EDIC Research Group (Writing Committee: Lachin JM, White N, Sun W, Cleary PA, Nathan D). Effect of intensive diabetes therapy on the progression of diabetes retinopathy in patients with type 1 diabetes: 18 years of follow-up in the DCCT/EDIC. *Diabetes* 2015; 64:631-42. PMC4303965.
DCCT/EDIC Research Group (Writing Committee: Nathan DM, Bebu I, Braffett BH, Orchard TJ, Cowie CC, Lopes-Virella M, Schutta M, Lachin JM). Risk factors for cardiovascular disease in type 1 diabetes. *Diabetes* 2016; 65:1370-76. PMC4829209.
DCCT/EDIC Research Group (Writing Committee: Gubitosi-Klug R, Lachin JM, Backlund JYC, Lorenzi GM, Brillon DJ, Orchard TJ). Intensive diabetes treatment and cardiovascular outcomes in type 1 diabetes: the DCCT/EDIC study 30-year follow-up. *Diabetes Care* 2016; 39:686-93. PMC4839174.
Demidenko E. *Mixed Models: Theory and Applications*. John Wiley & Sons, New York, 2006.
Dempster P, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 1977; 39:1-22.
Diabetes Epidemiology Research International Mortality Study Group. International evaluation of cause-specific mortality and IDDM. *Diabetes Care* 1991; 14:55–60.
Diggle, P. J., Heagerty P, Liang, K. Y. and Zeger, S. L. *Analysis of Longitudinal Data.* 2nd edition, New York: Oxford University Press, 2002.
Fine JP and Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* 1999; 94:496–509.
Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics* 1986; 42:845-54.
Fitzmaurice GM, Laird NM and Ware JH. Applied Longitudinal Analysis. 2nd Edition. John Wiley & Sons, New York, 2011.
Fleming TR and Harrington DP. Counting processes and survival analysis, John Wiley and Sons, Inc.: New York, 1991.
Food and Drug Administration (1998). International Conference on Harmonization: Guidance on statistical principles for clinical trials. *Federal Register* 1998; 63(179):49583–49598.
Frick H. A maximum linear test of normal means and its application to Lachin's data. *Comm Stat A* 1994; 23:1021-29.
Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; 71:431-44.

Gill TM, Gahbauer EA, Han L, Allore HG. Trajectories of disability in the last year of life. *N Engl J Med* 2010; 362:1173-80.
Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat* 1988; 16:1141-54.
Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; 61:92-105.
Hochberg Y and Tamhane AC. *Multiple Comparison Procedures.* John Wiley & Sons, New York, 1987.
Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Stat Med* 1997; 16:239–57.
Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statist* 1989;6:65-70.
Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med* 1998; 17:1623-34.
Ingel K, Jahn-Eimermacher A. Sample-size calculation and reestimation for a semiparametric analysis of recurrent event data taking robust standard errors into account. *Biom J* 2014; 56:631-48.
Inker LA, Schmid CH, Tighiouart H, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. *N Engl J Med* 2012; **367:** 20–29.
Jennrich RI and Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 1986; 42:805-20.
Kahn SE, Haffner SM, Heise MA, et al. Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. *N Engl J Med.* 2006;355:2427-2443.
Kalbfleisch JD and Lawless JF. The analysis of panel data under a Markov assumption. *J Am Stat Asso*c 1985; 80:863-71.
Kalbfleisch JD and Prentice RL. *The Statistical Analysis of Failure Time Data.* Second Edition. John Wiley & Sons, New York, 2002.
Jones BL, Nagin D. Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociological Methods and Research* 2007; 35(4):542-71.
Lachin JM. Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations. *Stat Med* 1992; 11:1151-70.
Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Clin Trials* 1999; 20:408-22.
Lachin, JM. *Biostatistical Methods: The Assessment of Relative Risks*. 2nd Edition. John Wiley and Sons; New York, 2011.
Lachin JM. Sample size and power for a logrank test and Cox proportional hazards model with multiple groups and strata, or a quantitative covariate with multiple strata. *Statistics in Medicine* 2013; 32:4413–4425.
Lachin JM. Power of the Mantel–Haenszel and other tests for discrete or grouped time-to-event data under a chained binomial model. *Statistics in Medicine* 2013; 32:220–229. DOI: 10.1002/sim.5480.
Lachin JM. Applications of the Wei-Lachin multivariate one-sided test for multiple outcomes on possibly different scales *PLoS ONE,* 2014; 9(10): e108784.
Lachin JM and Bebu I. Application of the Wei–Lachin multivariate one-directional test to multiple event-time outcomes. *Clinical Trials*, 2015, 12: 627-33.
Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*. 1986;42:507-519.
Lachin JM, Viberti G, Zinman B; for ADOPT Study Group. Renal function in type 2 diabetes with rosiglitazone, metformin, and glyburide monotherapy. *Clin J Am Soc Nephrol*. 2011;6:1032-1040.

Lachin JM and Wei LJ. Estimators and test in the analysis of nonindependent 2 x 2 tables with partially missing observations, *Biometrics* 1988; 44:513-28.
Laird NM and Oliver D. Covariance analysis of censored survival data using log-linear analysis techniques. *J Am Stat Assoc* 1981; 76:231-40.
Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38:963-74.
Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology* 2011; 22:575-81.
Lawless JF, Zhan M. Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Can J Stat* 1998; 26:549-65.
Liang K and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73:13-22.
Lin DY. Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of Amer Stat Assoc* 1991; 86:725-28.
Lin, D. Y. and Wei, L. J. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc* 1978; 84:1074-78.
Lin, D.Y., Wei L.J., Yang I. and Ying, Z. Semiparametric regression for the mean and rate functions of recurrent events. *J R. Stat Soc* 2000; 62:711–30.
Lin, DY, Ying, ZS. A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* 1993; 80:573-81.
Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993; 88:125–34.
Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annu Rev Public Health* 2000; 21:121-45.
Little RJA, Rubin DB. Statistical Analysis of Missing Data (second edition). New York NY: John Wiley and Sons, 2002.
Little RJ, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 1996; 52:98-111.
Lohr SL. Sampling: Design and Analysis, Pacific Grove, CA: Duxbury Press, 1999.
MacKinnon DP. Integrating mediators and moderators in research design. *Res Soc Work Pract* 2011; 21:675-81.
Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; 63:655-60.
Martinussen T, Scheike TH. *Dynamic regression models for survival data*, Springer, 2010.
McCullagh P and Nelder JA. *Generalized Linear Models*. 2nd ed. New York: Chapman & Hall, 1989.
Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. Handbook of missing data methodology, CRC Press, 2015.
Molenberghs G, Kenward MG. *Missing data in clinical studies*, Wiley, 2007.
Nathan DM, Zinman B, Cleary PA, Backlund JYC, Genuth S, Miller R, Orchard TJ, and the DCCT/EDIC Research Group. Modern-day clinical course of type 1 diabetes mellitus after 30 years duration. *Arch Intern Med* 2009; 169:1307-16. PMC2866072.
National Research Council of the National Academy of Science. *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, 2010.
Neter J, Kutner M, Nachtsheim C, Wasserman W. *Applied Linear Statistical Methods*. Irwin, 1996.
O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984; 40:1079-1087.
Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992; 48:951-59.

Pan W, Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data. *J Comput Graph Stat*, 1999; 8:109-20.
Parsons NR, Costa ML, Achten J, Stallard N. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Comput Stat Data Anal* 2009; 53:632-41.
Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *J Am Stat Assoc* 1991; 86:770-78.
Pepe MS and Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med* 1993; 12:737-51.
Pipper, CB, Ritz, C., Bisgaard, H. (2012). A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*. 61:315-326, 2012.
Pop-Busui R, Low PA, Waberski BH, Martin CL, Albers JW, Feldman EL, et al. Effects of prior intensive insulin therapy on cardiac autonomic nervous system function in type 1 diabetes mellitus: the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications study (DCCT/EDIC). *Circulation* 2009; 119:2886-93. PMC2757005.
Pregibon, D. Logistic regression diagnostics. *Ann Statist*, 1981; 9:705-24.
Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; 73:1-11.
Prentice RL, Williams BJ and Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika* 1981; 68:373-79.
Proust-Lima C, Sene M, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. *Stat Methods Med Res* 2014; 23:74-90.
Proust-Lima C, Sene M, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment PSA: a joint modeling approach. *Biostatistics* 2009; 10:535-49.
Ramlau-Hansen H. Smoothing counting process intensities by means of kernel functions. *Ann Stat* 1983; 11:453-66.
Ramlau-Hansen, H. The choice of a kernel function in the graduation of counting process intensities. *Scan Actuarial J* 1983;165-82.
Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70:41-55.
Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Statist Assoc* 1984; 79:516-24.
Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 1985; 39:33-38.
Rubin DB. Multiple *Imputation for Nonresponse in Surveys*. New York NY: John Wiley and Sons, 1987.
Rubin DB, Thomas N. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 1996; 52:249-64.
Schafer JL. *Analysis of Incomplete Multivariate Data*. New York NY: Chapman and Hall/CRC, 2000.
Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Stat Med* 1992; 11:1861-70.
Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statis Methods Med Res* 2013; 22:278-95.
Shih WJ. Sample size and power calculations for periodontal and other studies with clustered samples using the methods of generalized estimating equations. *Biom J* 1997; 39:899-908
Snedecor GW, Cochran WG. Statistical Methods (6th ed.) Ames, Iowa, Iowa State University Press, 1980.

Sparling YH, Younes N, Lachin JM and Bautista OM. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* 2006; 7:599-614.
Stevens RJ, Kothari V, Adler AI, Stratton IM and the United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes. *Clin Sci*. 2001;101:671-679.
Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci* 2010; 25:1-21.
Thall PF and Lachin JM. Analysis of recurrent events: Non-parametric methods for random-interval count data. *J Am Stat Assoc* 1988; 83:339-47.
Therneau TM and Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer-Verlag, 2000.
Therneau TM, Li H. Computing the Cox model for case cohort designs. *Lifetime Data Anal* 1999; 5:99-112.
Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997; 16:385-95.
Tobin J. Estimation of Relationships for Limited Dependent Variables. *Econometrica* 1958; 26:24-36.
Turnbull BW. The empirical distribution function with arbitrarily grouped, censored, and truncated data. *J R Stat Soc* 1976; 38:290-95.
Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014; 32:2380-85.
Vacek PM and Mickey RM et al. Application of a two-stage random effects model to longitudinal pulmonary function data from sarcoidosis patients. *Stat Med*, 1989; 8:189-200.
Van Buren S. Flexible imputation of missing data, Chapman & Hall/CRC, 2012.
VanderWeele TJ. A unification of mediation and interaction: A 4-way decomposition. *Epidemiology*. 2014; 25:749-61.
Vanderweele TJ, Explanation in causal inference: methods for mediation and interaction. Oxford University Press, 2015.
Wang L, McMahan CS, Hudgens MG, Qureshi ZP. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* 2016; 72:222-31.
Wei LJ and Lachin JM. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J Am Stat Assoc*, 1984; 79:653-61.
White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; 48:817-38.
White NH, Sun W, Cleary PA, Danis RP, Davis MD, Hainsworth DP, Hubbard L, Lachin JM, Nathan DM for the DCCT/EDIC Research Group. Prolonged Effect of Intensive Therapy on the Risk of Retinopathy Complications in Patients with Type 1 Diabetes Mellitus: 10 years after the Diabetes Control and Complications Trial, *Arch Ophthalmol* 2008; 126:1707-15. PMC2663518.
Whitehead J. Fitting Cox's regression model to survival data using GLIM, *J R Stat Soc* 1980; 29:268-75.
Willan AR, Lin DY, Cook RJ, Chen EB. Using inverse-weighting in cost-effectiveness analysis with censored data. *Stat Methods Med Res* 2002; 11:539–51.
Wu MC and Bailey KR. Analyzing changes in the presence of informative right censoring caused by death and withdrawal. *Stat Med*, 1988; 7:337-46.
Wu MC and Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*, 1989; 45:939-55.
Wu MC and Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; 44:175-88.

Wu MC, Hunsberger S, Zucker D. Testing for differences in changes in the presence of censoring: parametric and nonparametric methods. *Stat Med*, 1994; 113:635-46.
Younes N and Lachin JM. Link-based models for survival data with interval and continuous time censoring. *Biometrics* 1997; 53:1199-1211.
Zeger SL and Liang K. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42:121-30.
Zhao LZ, Claggett B, Tian L, Uno H, Pfeffer MA, Solomon SD, Trippa L, Wei LJ. On the restricted mean survival time curve in survival analysis. *Biometrics* 2016; 72:215-21.