



The **G**lycemia **R**eduction **A**pproaches in **D**iabetes: A Comparative Effectiveness **S**tudy (**GRADE Study**)

NCT01794143

**Statistical Analysis Plans for Manuscript Entitled
“Long term differences in metabolic status
among four initial treatments added to
metformin in early type 2 diabetes (OP1)” SAP**

Initial Statistical Analysis Plan

May 1, 2021

Final Statistical Analysis Plan

March 22, 2022

Summary of Changes to the Statistical Analysis Plan

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**

Statistical Analysis Plans for Manuscript Entitled “Long term differences in metabolic status among four initial treatments added to metformin in early type 2 diabetes (OP1)”

Table of Contents

INITIAL STATISTICAL ANALYSIS PLAN	2
This is the initial statistical analysis plan that was written and signed prior to locking the database and starting statistical analyses.	
FINAL STATISTICAL ANALYSIS PLAN	43
This is the final statistical analysis plan that describes the statistical analyses implemented in the manuscript.	
SUMMARY OF CHANGES TO STATISTICAL ANALYSIS PLAN	72
This describes the changes that were made to the initial statistical analysis plan prior to the final analyses that were included in the manuscript. These changes represent the differences between the initial statistical analysis plan and the final statistical analysis plan.	

Long term differences in metabolic status among four initial treatments added to metformin in early type 2 diabetes (OP1).

Table of Contents

GENERAL INFORMATION.....	2
APPROVALS.....	2
REVISION HISTORY.....	3
ABBREVIATIONS AND ACRONYMS.....	3
STUDY OBJECTIVES.....	3
Background and justification.....	3
Scientific objectives/questions.....	4
STATISTICAL METHODS AND DATASETS	5
Analysis Data Set Inclusion Criteria.....	5
Primary Variables to be Assessed.....	5
Statistical Analyses	7
Scientific Objective #1: Patient characteristics, retention, protocol completion, adherence by treatment group	7
Scientific Objective #2: Treatment effect on glycemic outcomes.....	11
Scientific Objective #3: Subgroup Analyses.....	18
Scientific Objective #4: Severe adverse events/targeted adverse events/side effects by treatment group.....	23
Scientific Objective #5: Mediation analyses.....	27
Scientific Objective #6: Sensitivity Analysis: Were estimated treatment effects affected by impact of the COVID-19 pandemic on the study data?	28
Scientific Objective #7: Sensitivity Analysis: Treatment effects among subset of GRADE data while on randomly assigned treatment (i.e., per-protocol analysis)	28
Scientific Objective #8: Sensitivity Analysis: Treatment effects if entire GRADE cohort had taken the assigned treatment according to study protocol during entire follow-up (inverse probability weighting analysis)	29
STATISTICAL CONSIDERATIONS.....	31
Rationale for Non-Standard Statistical Methodology.....	31
Inverse probability weighting (IPW) sensitivity analysis	31
Other statistical issues	32





Significance level	32
Intention-to-treat analyses.....	32
Definition of event times for glycemic outcomes (primary, secondary, and tertiary) ..	33
Checking the proportional hazards assumption for the Cox proportional hazards model.....	33
Adjustments for multiple pairwise comparisons among the treatment groups.....	33
Comparing each treatment to all other treatments combined	34
Adjustments for multiple comparisons for subgroup analyses	34
Calculation of confidence intervals adjusted for multiple comparisons based on the closed testing framework.....	35
APPENDIX A: Dataset Request.....	35
Table of Variables.....	35
Appendix B: Manuscript Figures Not Requiring Statistical Analysis	39
REFERENCES	40

GENERAL INFORMATION

Manuscript Title	Long term differences in metabolic status among four initial treatments added to metformin in early type 2 diabetes
GRADE paper number	OP1
Analysis Category	End of Study
Writing Group Chairs	David Nathan, John Lachin
Writing Group Members	David Nathan, John Lachin, Ashok Balasubramanyam, John Buse, Robert Cohen, Jill Crandall, Steven Kahn, Heidi Krause-Steinrauf, Mary Larkin, Neda Rasouli, Deborah Wexler
Target Journal	NEJM
Lead Statisticians	Naji Younes, Heidi Krause-Steinrauf, Nicole Butera

APPROVALS

	<i>Sign off on</i>	<i>Signature</i>
--	--------------------	------------------

David M. Nathan GRADE Co-PI and Writing Committee Co-Chair	Aims, outcomes, mock tables and figures	David Nathan  Digitally signed by David Nathan Date: 2021.05.06 19:54:13 -04'00'
John M. Lachin GRADE Co-PI and Writing Committee Co-Chair	Aims, outcomes, mock tables and figures, statistical methods	Move_Outlook2 016_Sent_Items  Digitally signed by Move_Outlook2016_Sent_Items Date: 2021.05.07 09:43:22 -04'00'
Naji Younes Supervisory Statistician	Outcomes, mock tables and figures, statistical methods	Naji Younes  Digitally signed by Naji Younes Date: 2021.05.11 15:48:40 -04'00'
Nicole Butera Analytic Statistician	Outcomes, mock tables and figures, statistical methods	Nicole Butera  Digitally signed by Nicole Butera Date: 2021.05.06 18:07:10 -04'00'

REVISION HISTORY

Version No.	Implemented by	Date	Reason
1	Nicole Butera	5/1/2021	Initial Version Approved

ABBREVIATIONS AND ACRONYMS

Abbreviation	Meaning
HbA1c	Glycated hemoglobin
T2DM	Type 2 diabetes

STUDY OBJECTIVES

Background and justification

Type 2 diabetes (T2DM) affects more than 30 million persons in the United States, with an incidence of 1.5 million new cases per year, and more than 400 million persons worldwide. The major human and economic costs associated with T2DM are related primarily to the development of long-term diabetes-specific complications, including retinopathy,

nephropathy, and neuropathy, and a 2-5 fold increased risk of non-specific cardiovascular disease (CVD). These long-term complications have been shown to be ameliorated in part by interventions that reduce chronic glycemia, as measured by glycated hemoglobin levels (HbA1c), and a target range of less than 7% (53 mmol/mol) has been established by consensus for most patients with T2DM. The estimated annual cost of diabetes in the US in 2017 was approximately \$327 billion dollars per year with an increasing fraction attributed to the cost of glucose-lowering medications.

Virtually all recommendations for the management of type 2 diabetes have included metformin as the first medication to be used. Unfortunately, choosing the second medication from the ever expanding list of glucose-lowering medications to add to metformin when monotherapy fails to achieve or maintain goal glycemia is problematic owing to the dearth of any long-term head-to-head comparator studies. The purpose of the Glycemia Reduction Approaches in Type 2 Diabetes: A Comparative Effectiveness (GRADE) Study was to examine the relative effectiveness of the four most commonly used glucose-lowering medications added to metformin to maintain goal glycemia. In this paper, we report the GRADE major glyceic outcomes. The accompanying paper reports the vascular outcomes and CVD risk factors associated with the four randomly assigned interventions.

Scientific objectives/questions

1. Summarize patient characteristics, retention, protocol completion, and adherence across the four treatment groups.
2. Do the primary, secondary, and/or tertiary glyceic outcomes differ by treatment group?
3. Do treatment effects on the primary, secondary, and/or tertiary glyceic outcomes vary by the following pre-specified baseline subgroups: race/ethnicity, gender, age, diabetes duration, BMI, HbA1c?
4. Do severe adverse events/targeted adverse events/side effects differ by treatment group?
5. Are treatment effects on the primary, secondary, and/or tertiary glyceic outcomes mediated by other factors?
6. Sensitivity Analysis 1: Were the estimated treatment effects on the primary, secondary, and/or tertiary glyceic outcomes affected by the impact of the COVID-19 pandemic on the study data?
7. Sensitivity Analysis 2: What were the treatment effects on the primary, secondary, and/or tertiary glyceic outcomes among the subset of the GRADE data while on the randomly assigned treatment (i.e., per-protocol analysis)?
8. Sensitivity Analysis 3: What would the treatment effects on the primary, secondary, and/or tertiary glyceic outcomes have been if the entire GRADE cohort had taken the assigned treatment according to study protocol during the entire follow-up period?

STATISTICAL METHODS AND DATASETS

Analysis Data Set Inclusion Criteria

Full analysis set: All available follow-up data from all randomized participants (n=5047).

For scientific objective #7, will use the *per-protocol* data set:

- The subset of participants who meet both of the following criteria:
 - Took at least one dose of the assigned therapy
 - Completed at least one outcome assessment visit
- The subset of participant data that meets the following criteria:
 - Data up to the end of study for patients who do not discontinue the assigned drug regimen for greater than 4 weeks during the study and/or initiate the use of non-study diabetes drug(s).
 - Data prior to the first discontinuation of the assigned drug regimen for greater than 4 weeks (28 days), for patients who so discontinued. A subject is considered to have discontinued from the study regimen (i.e., the assigned medications according to the study protocol) if the subject stops taking at least one of the study medications called for under the regimen (e.g., a subject who fails to start glargine after reaching the secondary outcome is considered to have discontinued the study regimen). In particular, note that data following discontinuation of the randomly assigned medication after reaching the tertiary outcome would be excluded from the per-protocol dataset.
 - Data prior to initiation of use of non-study diabetes drug(s).
 - Data up to the time of withdrawal from the study.

Primary Variables to be Assessed

Treatment assignment: Glimepiride (Sulfonylurea), Liraglutide (GLP-1 RA), Sitagliptin (DPP 4-inhibitor), Glargine (Insulin)

HbA1c during follow-up

Glycemic outcomes (*definitions from study protocol*):

- **Primary glycemic outcome:** Time to an initial HbA1c $\geq 7\%$, subsequently confirmed at the next quarterly visit. If the initially observed HbA1c is $> 9\%$, then the confirmation value will be performed within 3 to 6 weeks. If the initial HbA1c and confirmation value 3 to 6 weeks later are both $> 9\%$, the primary and secondary outcomes will have been reached. If the initial HbA1c is $> 9\%$ and the confirmation value 3 to 6 weeks later is $\leq 9\%$, the participant will resume his/her usual schedule of quarterly HbA1c monitoring. If the HbA1c at the next quarterly visit is $\geq 7\%$, then the

primary outcome will have been reached. The primary outcome can only be reached after a minimum of 6 months of therapy, unless the HbA1c at 3 months is > 9% and is higher for the confirmation HbA1c 3-6 weeks later, in which case the primary and secondary outcomes will have been met at 3 months.

- **Secondary glyceimic outcome:** Time to an HbA1c > 7.5% after having reached the primary outcome, subsequently confirmed at the next quarterly visit. The primary and secondary outcomes may be reached simultaneously if the initial value and the confirmation are both > 7.5%.
- **Tertiary glyceimic outcome:** Time to an HbA1c > 7.5% after having confirmed the secondary outcome (at which point the participant should have started basal insulin based on study protocol), subsequently confirmed at the next quarterly visit. Note that following the intention-to-treat framework, the main analyses for this paper will define the tertiary outcome irrespective of whether participants actually start basal insulin following a secondary outcome according to study protocol (*based on decision made during Writing Group call on 11/24/2020*).

Study compliance variables:

- **Visit adherence:** $100\% * (\text{number of study visits attended}) / (\text{expected number of study visits according to study protocol})$
- **Duration of follow-up:** Date of last study contact minus randomization date
- **Permanent discontinuation of metformin:** The participant reports not taking metformin at all subsequent study visits
- **Permanent discontinuation of study treatment regimen:** Stopping at least one of the study medications called for based on the study protocol. The participant reports not taking the medication at all subsequent study visits.
- **Temporary discontinuation of study treatment regimen:** Stopping at least one of the study medications called for based on the study protocol. The assigned study treatment regimen has been stopped for greater than 4 weeks (28 days).
- **Off-study use of glucose-lowering medication:** Overall, and specifically for off-study medications in the following classes: sulfonylurea, DPP 4-inhibitor, GLP-1 RA, insulin, SGLT-2 inhibitor, thiazolidinedione, other

Adverse events and side effects:

- Mortality
- Any adverse event (targeted event or event resulting in hospitalization overnight or \geq 24 hours)
- Serious adverse event
- Hospitalization overnight or \geq 24 hours

- Severe or major hypoglycemia
 - Severe hypoglycemia requires 3rd party assistance
 - Major hypoglycemia is a severe episode that results in loss of consciousness and/or seizure
 - Severe hypoglycemia that results in injury to the participant or others (e.g. motor vehicle accident in which the participant was the driver)
- Weight gain \geq 10% higher than at randomization
- Gastrointestinal symptoms (nausea, vomiting, diarrhea, stomach pain/bloating)
- Lactic acidosis
- Pancreatitis
- Acute metabolic decompensation (diabetic ketoacidosis, HHS)
- Gallstone disease (cholelithiasis, cholecystitis)
- Cancer (thyroid, pancreatic, other)

NOTE: a table listing of all variables is provided in the Appendix hereto.

Statistical Analyses

Scientific Objective #1: Patient characteristics, retention, protocol completion, adherence by treatment group

Supplemental Table 1. Selected baseline characteristics related to the glycemic outcomes.

Continuous row variables: Mean (SD) for row variable overall and stratified by treatment group, unless otherwise specified.

Binary/categorical row variables: n (%) for row variable overall and stratified by treatment group.

	All	Insulin Glargine	Glimepiride	Liraglutide	Sitagliptin
n (%)	xxx	xxx	xxx	xxx	xxx
Age					
<45 years	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
45-59 years	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)

≥60 years	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Women (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Race					
Am Ind/Alaska Native	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Asian	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Hawaiian/Pacific Isl	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Black or African-American	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
White	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Other/multiple	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Unknown/not reported	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Ethnicity	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Duration of diabetes (years), Mean (SD)	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Duration of diabetes (years), Median (IQR)	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]
Baseline metformin dose					
1000	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
1500	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
2000	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
BMI (kg/m ²)	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
HbA1c	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x

Table 1. Retention, protocol completion and adherence comparing the treatment groups during the entire GRADE study period.

Discontinuation of assigned treatment off-protocol refers to stopping at least one of the study medications called for based on the study protocol (e.g., a participant who fails to

start glargine after reaching the secondary outcome is considered to have discontinued the assigned treatment, whereas stopping the randomized medication due to reaching the tertiary outcome is not considered to have discontinued). A participant will be considered to **permanently discontinue** a study medication if the participant reports not taking the medication at all subsequent study visits. A participant will be considered to **temporarily discontinue** the assigned treatment if the assigned study treatment regimen has been stopped for greater than 4 weeks (28 days).

Continuous row variables: Mean (SD) for row variable overall and stratified by treatment group, unless otherwise specified.

Binary/categorical row variables: n (%) for row variable overall and stratified by treatment group.

	All	Insulin Glargine	Glimepiride	Liraglutide	Sitagliptin
N	xxx	xxx	xxx	xxx	xxx
Attended close-out study visit (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Visit adherence (%) ¹	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Mean duration of follow-up (years) ²	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Median duration of follow-up (years) ²	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]
Number (%) of participants who permanently discontinued metformin	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Number (%) that permanently discontinued assigned study treatment regimen off-protocol ³	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Within 1st year post-randomization	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
1 - 2 years post-randomization	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
2+ years post-randomization	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Number (%) that temporarily discontinued assigned study treatment regimen off-protocol ⁴	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)

Time (years) on assigned study treatment regimen per protocol ³	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
% of study time on assigned study treatment regimen per protocol ^{3,5}	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Use of non-study, off-protocol glucose-lowering medications (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Sulfonylurea (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
DPP 4-inhibitor (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
GLP-1 RA (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Insulin (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
SGLT-2 inhibitor (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Thiazolidinedione (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Other (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)

¹ Visit adherence = 100% * (number of study visits attended) / (expected number of study visits according to the study protocol), calculated for each individual

² Duration of follow-up = date of last study contact – date of randomization

³ Only includes treatment discontinuation that was not consistent with the study protocol. Specifically, this does not include discontinuation of the randomized medication due to reaching the tertiary outcome, as required by study protocol.

⁴ Participants were considered to have temporarily discontinued the assigned treatment regimen if the study medication(s) were discontinued for a minimum of 4 weeks.

⁵ Percent of time from randomization to date of last study contact calculated for each individual

Scientific Objective #2: Treatment effect on glycemc outcomes

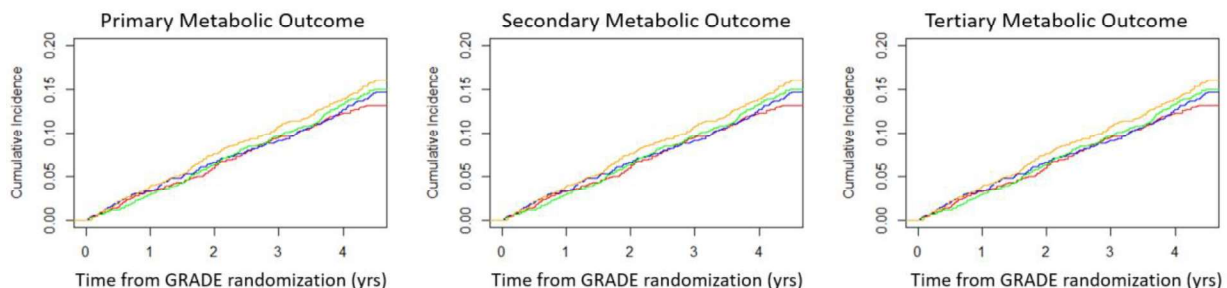
Figure 1. Cumulative incidence of (a) primary, (b) secondary and (c) tertiary glycemc outcomes by treatment group. Mean (d) HbA1c, (e) fasting plasma glucose levels and (f) weight over study time.

A 2x3-panel figure.

The 3 panels in the top row display the cumulative incidence for the primary, secondary, and tertiary outcomes (from left to right) over time. Each panel includes 4 lines, one for the cumulative incidence within each treatment group. The cumulative incidence by treatment group will be estimated using a Kaplan-Meier estimator. The total number at risk at each year will be provided below each panel. The time axis will represent the time since GRADE randomization. The maximum value for the time axis will be selected as the last time when the total number at risk is ≥ 200 for the primary outcome. The y-axis limits will be selected to be the same for the panels for the primary, secondary, and tertiary outcomes, per journal requirements.

The 3 panels in the bottom row display the mean values of HbA1c (%), fasting glucose (mg/dL), and weight (kg) (from left to right) over time. HbA1c and weight were collected at each quarterly visit, and so means will be displayed for every 3 months. Fasting glucose was collected at baseline, 1 year, 3 years, and 5 years post-randomization, and therefore means will be displayed for these time-points. 95% confidence bands for the longitudinal means will be graphed, based on a simple repeated measures model for the longitudinal means without covariates. Each panel includes 4 lines, one for longitudinal means with each treatment group. For consistency, the same time axis will be used for these panels as for the top panels displaying cumulative incidence of the primary, secondary, and tertiary glycemc outcomes. The number of participants at each year will be provided below each panel. In addition, graphs will be presented for the kernel-smoothed distributions of HbA1c, fasting glucose, and weight by treatment group at 1 year and 3 years post-randomization.

A simple mocked-up version of this figure using simulated data is displayed below.



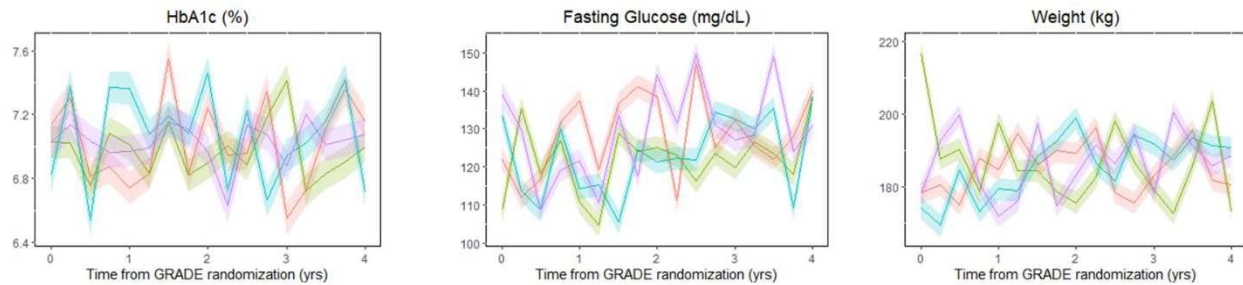
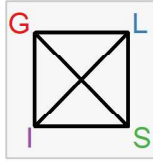


Table 2. Numbers of subjects reaching primary, secondary and tertiary glycemc outcomes by treatment group, with crude rates, pairwise hazard ratios, and hazard ratios compared to all other treatments combined.




For this table, the following statistics will be calculated for the primary, secondary, and tertiary glycemc outcomes, both overall and stratified by treatment group:



- The number of events and percent of the GRADE cohort with the outcome.
- Crude rate per 100 person-years (SE). This will be calculated as $100 \times (\text{observed number of events}) / (\text{total time at risk})$, where the total time at risk is the sum of the time since randomization to the event (or to the censoring time for those without an event) across participants.
- Pairwise hazard ratios (SE). A Cox proportional hazards model will be fit for the outcome with treatment group as a predictor. For the purposes of this Cox model, the event times and censoring times will be calculated as time since randomization to the event or censoring respectively. Hazard ratios and standard errors for each pairwise comparison of the treatment groups will be estimated from the Cox model. All Wald-type tests, standard errors and confidence intervals will be estimated using the robust Lin-Wei (1989) information sandwich estimator to ensure valid inferences even if the proportional hazards assumption does not apply. A joint test for differences in the hazards among any of the treatment groups will be conducted. If that joint test is significant, then pairwise log-rank tests will be conducted to test for all pairwise differences. There are a total of 6 possible pairwise comparisons among the 4 treatment groups, and therefore these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). If the joint test for differences among any of the treatment groups is significant, then the results from the pairwise testing will be visualized using the following graphic, where each corner of the box represents one of the four treatments (G = Glimperide, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly; dotted lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and solid lines indicate $p \leq 0.001$.

Pairwise Tests



- Hazard ratio compared to all other treatments combined (SE) (Lachin and Bebu, 2020). A Cox proportional hazards model will be fit for the outcome with treatment group as a predictor. For the purposes of this Cox model, the event times and censoring times will be calculated as time since randomization to the event or censoring respectively. For treatment a , the hazard ratio compared to all other treatments combined will be estimated as the average of the estimated hazard ratios comparing each of the other treatments to treatment a . Since there are 4 treatment groups, there would be a total of 4 tests comparing each treatment to all others combined, and therefore these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document).
- Pairwise RMST ratios (SE). A log-linear model will be fit for the restricted mean survival time (RMST) up to $\tau = 4$ years using inverse probability of censoring weighting (IPCW) (Tian et al, 2014). RMST ratios and standard errors for each pairwise comparison of the treatment groups will be estimated from this model. The same testing procedure for the pairwise comparisons will be used as for testing pairwise hazard ratios above.
- RMST ratio compared to all other treatments combined (SE). A log-linear model will be fit for the restricted mean survival time (RMST) up to $\tau = 4$ using inverse probability of censoring weighting (IPCW) (Tian et al, 2014). For treatment a , the RMST ratio compared to all other treatments combined will be estimated as the average of the estimated RMST ratios comparing each of the other treatments to treatment a . The same testing procedure will be used as for testing hazard ratios compared to all other treatments combined above.
- According to the study protocol, participants should add glargine insulin to their treatment regimen after reaching the secondary outcome. Therefore, this table will also report the number and percent of participants who actually start treatment with glargine insulin following the secondary outcome. The percent will be calculated as $100\% * (\text{number of participants who started glargine insulin}) / (\text{number of participants who reached the secondary outcome})$.

	Total (N=5047)	Glimepiride (G) (N=XXXX)	Liraglutide (L) (N=XXXX)	Sitagliptin (S) (N=XXXX)	Insulin Glargine (I) (N=XXXX)	Pairwise Treatment Comparisons ¹
Primary Metabolic Outcome						
n (%)						
Crude rate per 100 person- years (SE)						
Pairwise hazard ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 
Hazard ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
Pairwise RMST ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 
RMST ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
Secondary Metabolic Outcome						
n (%)						
Crude rate per 100 person- years (SE)						
Pairwise hazard ratios (SE)						
Glimepiride (SE) Liraglutide (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 

Sitagliptin (SE)						
Hazard ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
Pairwise RMST ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 
RMST ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
Number (%) starting on insulin after reaching secondary outcome	--				--	
Tertiary Metabolic Outcome						
n (%)						
Crude rate per 100 person-years (SE)	--					
Pairwise hazard ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 
Hazard ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
Pairwise RMST ratios (SE)						

Glimepiride (SE)	--	--	--			Pairwise Tests
Liraglutide (SE)	--	--	--	--		
Sitagliptin (SE)						
RMST ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=

* $p \leq 0.05$ (from test that hazard ratio equals 1)

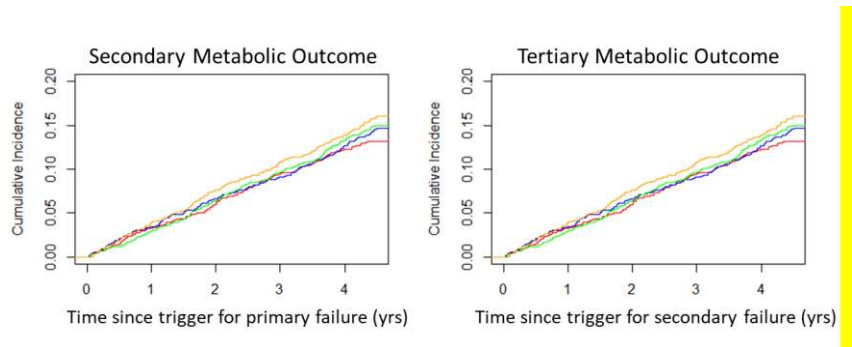
** $p \leq 0.01$ (from test that hazard ratio equals 1)

*** $p \leq 0.001$ (from test that hazard ratio equals 1)

¹ Boxes in this column graphically display the results of testing pairwise treatment effects. Each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly. Solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

Supplementary Figure 3. Cumulative incidence of (a) secondary outcome relative to time since primary outcome and (b) tertiary outcome relative to time since secondary outcome, by treatment group.

This will be a 2-panel figure, displaying the cumulative incidence for the secondary and tertiary outcomes (from left to right) over time, where the time axis will represent the time since the trigger HbA1c value for the primary or secondary outcome respectively. The panels in this figure will be similar to the top panels in Figure 1, except that the time variable corresponds to time since the trigger for the primary or secondary outcome, instead of time since randomization. A simple mocked-up version of this figure using simulated data is displayed below.




Supplementary Table 2. Crude rates, pairwise hazard ratios, and hazard ratios compared to all other treatments combined for (a) secondary outcome relative to time since primary outcome and (b) tertiary outcome relative to time since secondary outcome, by treatment group

This table displays similar statistics as Table 2 for the secondary and tertiary glycemic outcomes, except that for this table the event times are calculated as the time since the

trigger HbA1c value for the primary or secondary outcome respectively to the event (i.e., instead of as the time since randomization to the outcome). In particular, the following statistics will be calculated for the secondary and tertiary glycemc outcomes, stratified by treatment group:

- Crude rate per 100 person-years (SE). This will be calculated as $100 \times (\text{observed number of events}) / (\text{total time at risk})$, where the total time at risk is the sum of the time since the trigger HbA1c value for the previous outcome to the event (or to the censoring time for those without an event) across participants.
- Pairwise hazard ratios (SE). These will be estimated in a similar way to Table 2, except that the event times and censoring times will be calculated as time since the trigger HbA1c value for the previous outcome to the event or censoring respectively.
- Hazard ratio compared to all other treatments combined (SE). These will be estimated in a similar way to Table 2, except that the event times and censoring times will be calculated as time since the trigger HbA1c value for the previous outcome to the event or censoring respectively.
- Pairwise RMST ratios (SE). These will be estimated in a similar way to Table 2, except that the event times and censoring times will be calculated as time since the trigger HbA1c value for the previous outcome to the event or censoring respectively.
- RMST ratio compared to all other treatments combined (SE). These will be estimated in a similar way to Table 2, except that the event times and censoring times will be calculated as time since the trigger HbA1c value for the previous outcome to the event or censoring respectively.

	Glimepiride (G) (N=XXXX)	Liraglutide (L) (N=XXXX)	Sitagliptin (S) (N=XXXX)	Insulin Glargine (I) (N=XXXX)	Pairwise Treatment Comparisons ¹
Secondary Metabolic Outcome (Relative to time since trigger for primary failure)					
Crude rate per 100 person-years (SE)					
Pairwise hazard ratios (SE)					
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- --	--		Pairwise Tests 
Hazard ratio compared to all other treatments combined (SE)					G: p= L: p= S: p= I: p=
Pairwise RMST ratios (SE)					

Glimepiride (SE)	--	--	--		
Liraglutide (SE)	--	--	--		
Sitagliptin (SE)	--	--	--		
RMST ratio compared to all other treatments combined (SE)					G: p= L: p= S: p= I: p=
Tertiary Metabolic Outcome (Relative to time since trigger for secondary failure)					
Crude rate per 100 person-years (SE)					
Pairwise hazard ratios (SE)					
Glimepiride (SE)	--	--	--		
Liraglutide (SE)	--	--	--		
Sitagliptin (SE)	--	--	--		
Hazard ratio compared to all other treatments combined (SE)					G: p= L: p= S: p= I: p=
Pairwise RMST ratios (SE)					
Glimepiride (SE)	--	--	--		
Liraglutide (SE)	--	--	--		
Sitagliptin (SE)	--	--	--		
RMST ratio compared to all other treatments combined (SE)					G: p= L: p= S: p= I: p=

* $p \leq 0.05$ (from test that hazard ratio equals 1)

** $p \leq 0.01$ (from test that hazard ratio equals 1)

*** $p \leq 0.001$ (from test that hazard ratio equals 1)

¹ Boxes in this column graphically display the results of testing pairwise treatment effects. Each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly. Solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

Scientific Objective #3: Subgroup Analyses

Subgroup analyses of the treatment effects for the primary, secondary, and tertiary glycemic outcomes within subgroups based on the following baseline variables: race/ethnicity, sex, age, diabetes duration, body mass index, and HbA1c. The subgroups for each baseline variable are defined as the following:

- Race/ethnicity: Non-Hispanic white, non-Hispanic black, Hispanic white, and other (the “other” category includes all race/ethnicity categories other than non-Hispanic white, non-Hispanic black, or Hispanic white, and includes participants who specified other or unknown race)
- Sex: Male, female
- Age: <45 years, 45-59 years, 60+ years
- Diabetes duration, body mass index, HbA1c: Sample tertiles

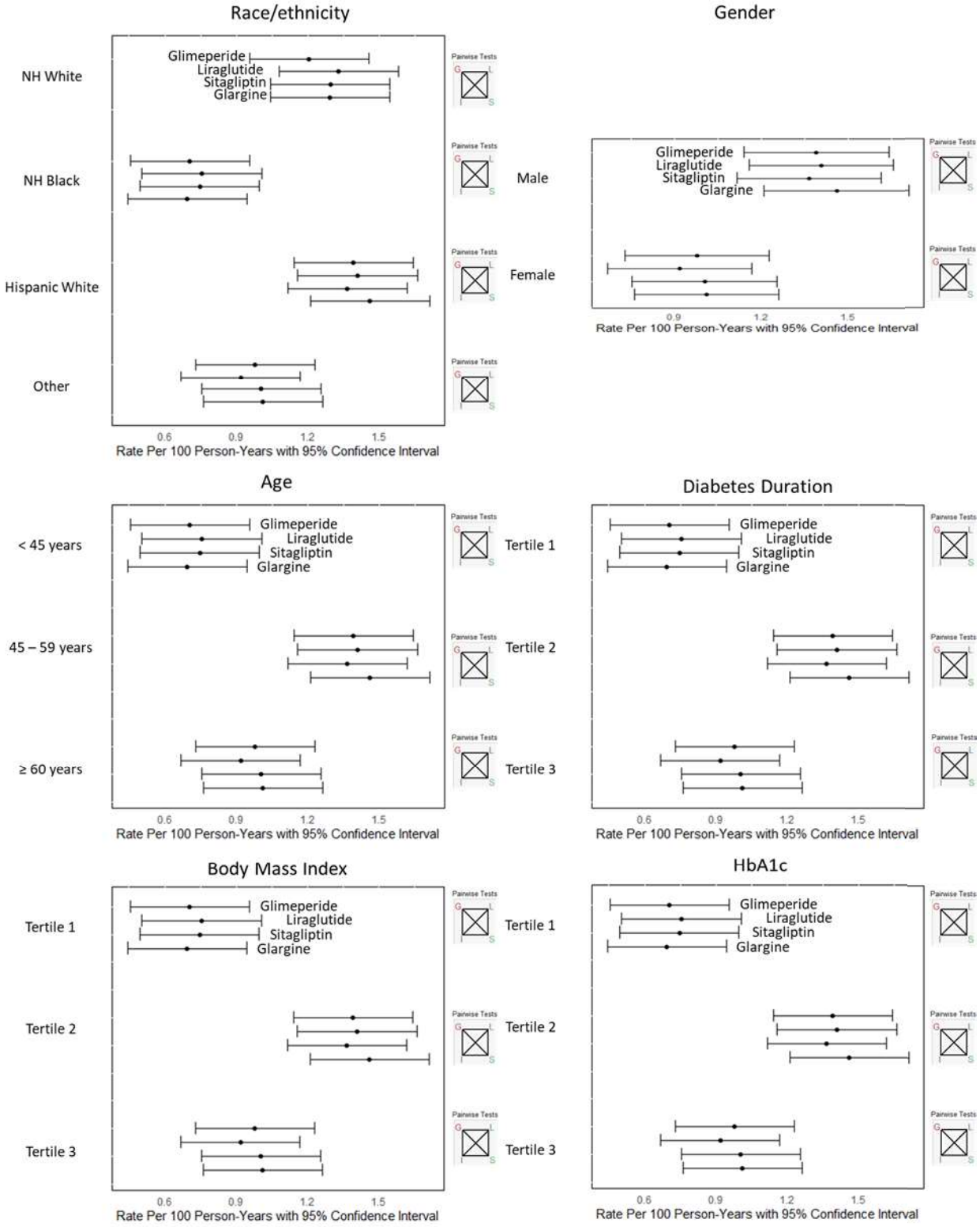
Figure 2. Analyses for primary outcome within protocol pre-specified baseline subgroups

A 3x2-panel figure. There is a separate panel for each of the baseline subgroup variables. Tests of pairwise treatment comparisons within each subgroup will be assessed. Since there are a total of 6 possible pairwise comparisons within each subgroup, these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). The results from the pairwise testing will be visualized using a graphic, where each corner of the box represents one of the four treatments (G = Glimpiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly; solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$. For subgroups that are determined not to have heterogeneous treatment effects, treatment effects will not be tested within subgroup, and so this graphic will be omitted.

Each panel will display the crude rates per 100 person-years (with 95% confidence intervals) of the primary outcome for each treatment group within each subgroup of the baseline variable. The crude rates will be calculated as $100 \times (\text{observed number of events}) / (\text{total time at risk})$, where the total time at risk is the sum of the time since randomization to the event (or to the censoring time for those without an event) across participants.

A simple mocked-up version of this figure using simulated data is displayed below.


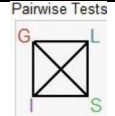
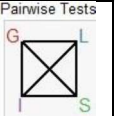

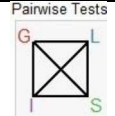
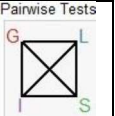


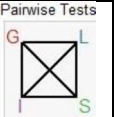
A similar figure will be produced for the secondary outcome and the tertiary outcome also.



Supplementary Table 3. Analyses for primary, secondary, and tertiary outcome within protocol pre-specified baseline subgroups

The following statistics will be calculated for treatment effects on the primary, secondary, and tertiary glyceimic outcomes within baseline subgroups:

- The number of participants in each treatment group within each subgroup.
- The number of events for each glyceimic outcome in each treatment group within each subgroup.
- Crude rate per 100 person-years (with 95% confidence intervals) of each glyceimic outcome in each treatment group within each subgroup. The crude rates will be calculated as $100 * (\text{observed number of events}) / (\text{total time at risk})$, where the total time at risk is the sum of the time since randomization to the event (or to the censoring time for those without an event) across participants.
- P-value from overall test of homogeneity of treatment effect across each baseline subgroup variable. For quantitative factors (i.e., age, diabetes duration, BMI, HbA1c), this p-value will be based on a test of homogeneity of treatment effect across the continuous quantitative variable (i.e., a test of covariate by group interaction, not based on the categorized subgroups).
- Hierarchical closed testing of subgroup by group interaction will be used to identify subgroups within which some heterogeneity may exist, and within each such subgroup the treatment groups will be compared using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document; Lachin, et al. 2019). Tests of pairwise treatment comparisons within a subgroup will be visualized in the same way as the tests of pairwise treatment comparisons within subgroups in Figure 2.

Subgroup (p) ¹	Treatment Group	Outcome Event Rates ²									
		Primary Outcome ³				Secondary Outcome ³			Tertiary Outcome ³		
Total		N	N _e	Rate(C I)	Tests ³	N _e	Rate(C I)	Tests ³	N _e	Rate(C I)	Tests ³
	Glimepiride (G)	N	n	r(l, ul)	Pairwise Tests 	n	r(l, ul)	Pairwise Tests 	n	r(l, ul)	Pairwise Tests 
	Liraglutide (L)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
	Sitagliptin (S)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
	Insulin Glargine (I)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
Sex (p=0.xx)											
Male	Glimepiride (G)	N	n	r(l, ul)	Pairwise Tests 	n	r(l, ul)	Pairwise Tests 	n	r(l, ul)	Pairwise Tests 
	Liraglutide (L)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
	Sitagliptin (S)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
	Insulin Glargine (I)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
Female	Glimepiride (G)	N	n	r(l, ul)	Pairwise Tests 	n	r(l, ul)	Pairwise Tests 	n	r(l, ul)	Pairwise Tests 
	Liraglutide (L)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
	Sitagliptin (S)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	
	Insulin Glargine (I)	N	n	r(l, ul)		n	r(l, ul)		n	r(l, ul)	

Tertile 2	Glimepiride (G)	N	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests
	Liraglutide (L)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Sitagliptin (S)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Insulin Glargine (I)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
Tertile 3	Glimepiride (G)	N	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests
	Liraglutide (L)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Sitagliptin (S)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Insulin Glargine (I)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
Diabetes Duration (p=0.xx)											
Tertile 1 yrs	Glimepiride (G)	N	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests
	Liraglutide (L)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Sitagliptin (S)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Insulin Glargine (I)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
Tertile 2	Glimepiride (G)	N	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests
	Liraglutide (L)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Sitagliptin (S)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Insulin Glargine (I)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
Tertile 3	Glimepiride (G)	N	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests 	n	r(II, ul)	Pairwise Tests
	Liraglutide (L)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Sitagliptin (S)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	
	Insulin Glargine (I)	N	n	r(II, ul)		n	r(II, ul)		n	r(II, ul)	

¹Subgroup levels (p-value for test of homogeneity of treatment effect across subgroup levels)

²Event rates are crude rates expressed as number of events per 100 patient-years of followup

³Boxes in this column graphically display the results of testing pairwise treatment effects. Each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly. Solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

⁴The “other” race/ethnicity category includes all race/ethnicity categories other than non-Hispanic white, non-Hispanic black, or Hispanic white, and includes participants who specified other or unknown race.

Scientific Objective #4: Severe adverse events/targeted adverse events/side effects by treatment group

Table 3. Overall incidence and rate of adverse events and side effects with comparison of treatment groups



This table displays the number and percent of participants who experienced each type of adverse event/side effect, and the crude rate (SE) of the adverse event/side effect, stratified by treatment group. The crude rates per 100 person-years will be calculated as $100 * (\text{observed number of events}) / (\text{total time at risk})$.







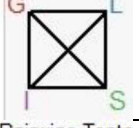

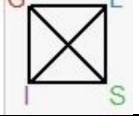
Tests of all pairwise treatment comparisons for each adverse event/side effect will be assessed, based on a Poisson regression of the number of events with the log of exposure time as an offset and treatment as the only covariate. Since there are a total of 6 possible pairwise comparisons for each adverse event/side effect, these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). The results from the pairwise testing will be





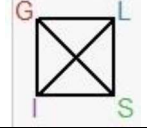



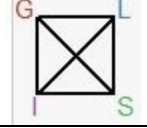
visualized using a graphic, where each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly; solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

The following adverse events/side effects will be included in this table:

- Mortality
- Any adverse event (targeted event or event resulting in hospitalization ≥ 24 hours)
- Serious adverse event
- Hospitalization overnight or ≥ 24 hours
- Severe or major hypoglycemia
- Weight gain $\geq 10\%$ higher than at randomization
- Gastrointestinal symptoms (nausea, vomiting, diarrhea, stomach pain/bloating)
- Lactic acidosis
- Pancreatitis
- Acute metabolic decompensation (diabetic ketoacidosis, HHS)
- Gallstone disease (cholecystitis and cholelithiasis)
- Thyroid cancer (all, medullary)
- Pancreatic cancer
- Other cancer

Adverse Event	Glimepiride (G) (N=XXXX)		Liraglutide (L) (N=XXXX)		Sitagliptin (S) (N=XXXX)		Insulin Glargine (I) (N=XXXX)		Pairwise Treatment Comparisons ¹
	n (%)	Rate (SE) ²	n (%)	Rate (SE) ²	n (%)	Rate (SE) ²	n (%)	Rate (SE) ²	
Mortality ³									Pairwise Tests 
Any adverse event (targeted event or event resulting in hospitalization ≥ 24 hours)									Pairwise Tests 

Serious adverse event									Pairwise Tests 
Hospitalization overnight or \geq 24 hours									
Severe hypoglycemia ^{3,4}									Pairwise Tests 
Major hypoglycemia ^{3,5}									Pairwise Tests 
Weight gain \geq 10% higher than at randomization									Pairwise Tests 
Gastrointestinal symptoms									Pairwise Tests 
Nausea									Pairwise Tests 
Vomiting									Pairwise Tests 
Diarrhea									Pairwise Tests 
Stomach pain / bloating									Pairwise Tests 

Lactic acidosis ³									Pairwise Tests 
Pancreatitis ³									Pairwise Tests 
Acute metabolic decompensation									Pairwise Tests 
Diabetic ketoacidosis									Pairwise Tests 
Hyperosmolar Hyperglycemic Syndrome									Pairwise Tests 
Gallstone disease (cholecystitis and cholelithiasis)									Pairwise Tests 
Thyroid cancer ³ All Medullary									Pairwise Tests 
Pancreatic cancer ³									Pairwise Tests 
Other cancer ³									Pairwise Tests 

¹ Boxes in this column graphically display the results of testing pairwise treatment effects. Each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly. Solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

² Event rate per 100 person-years. HHS- hyperglycemic hyperosmolar syndrome.

³ Adjudicated events

⁴ Defined as an episode requiring third-party assistance to treat

⁵ Defined as an episode resulting in coma/seizure

Scientific Objective #5: Mediation analyses

Supplementary Table 4. Treatment group differences without and with adjustment for potential mediators

Mediation analyses will be conducted to estimate the proportion of treatment effects on the glycemic outcomes that are explained by weight as a mediator. This analysis will follow Baron and Kenny's mediation paradigm (Baron & Kenny, 1986). First, an unadjusted model for the outcome by treatment group will be fit, and the treatment effect of each treatment vs. all others will be estimated from this model (θ_{0k} for treatment k). Then, a second model will be fit for the treatment effect on the outcome adjusted for the current value of weight (i.e., weight as a time-varying covariate), and the treatment effect of each treatment vs. all others will also be estimated from this model (θ_{1k} for treatment k). Finally, the percent mediation of the treatment effect by weight for each treatment will be calculated as the relative change in the treatment effect in a model adjusted for weight as a mediator relative to an unadjusted model (i.e., $\frac{\theta_{0k}-\theta_{1k}}{\theta_{0k}} * 100\% = (1 - \theta_{1k}/\theta_{0k}) * 100\%$).

Since the proportional hazards assumption is not preserved under marginalization (i.e., the proportional hazards assumption cannot hold for both the unadjusted model and the model adjusted for weight as a mediator; Gail et al, 1984), standard errors will be estimated using the robust Lin-Wei (1989) information sandwich estimator to ensure valid inferences when the proportional hazards assumption does not apply.

Outcome	Model Adjustment	Glimepiride vs others		Liraglutide vs others		Sitagliptin vs others		Insulin Glargine vs others	
		HR (SE) ²	% Med ^{1,2}	HR (SE) ²	% Med ^{1,2}	HR (SE) ²	% Med ^{1,2}	HR (SE) ²	% Med ^{1,2}
Primary	None	$\theta_{01} (\sigma_{01})$	--	$\theta_{02} (\sigma_{02})$	-	$\theta_{03} (\sigma_{03})$	-	$\theta_{04} (\sigma_{04})$	-
	Weight	$\theta_{11} (\sigma_{11})$	$\phi_{11} (\sigma_{11})$	$\theta_{12} (\sigma_{12})$	$\phi_{12} (\sigma_{12})$	$\theta_{13} (\sigma_{13})$	$\phi_{13} (\sigma_{13})$	$\theta_{14} (\sigma_{14})$	$\phi_{14} (\sigma_{14})$
Secondary	None		--		--		--		--
	Weight								
Tertiary	None		--		--		--		--
	Weight								

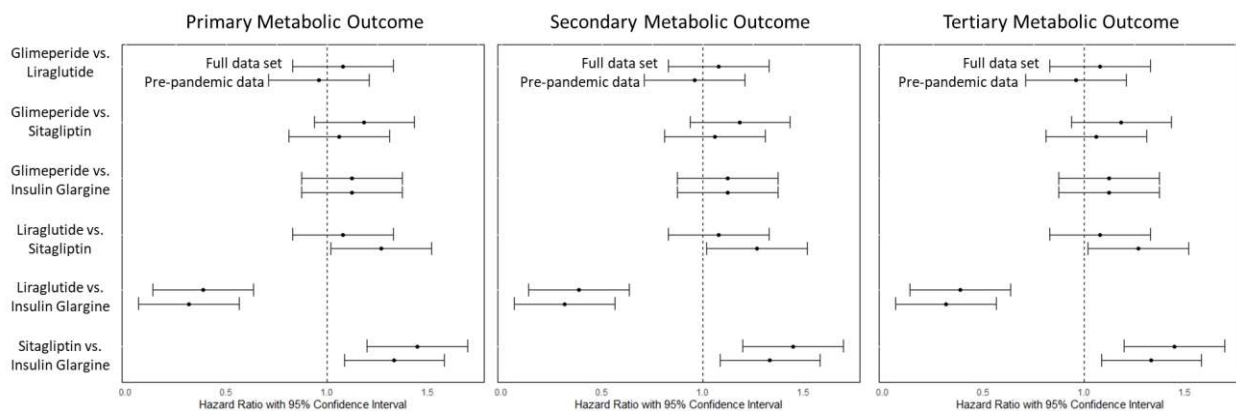
¹Percent mediation was calculated as the relative change in the hazard ratio for treatment group in a model adjusted for the mediator relative to an unadjusted model. For example, ϕ_{11} is the estimate of the percent change in the hazard ratio (Insulin Glargine vs others) in the model adjusted for Mediator 1 relative to the unadjusted model and is calculated as $(\theta_{11}/\theta_{01} - 1) * 100$ where the HR estimates (θ 's) are calculated from the appropriate contrasts of the model coefficients from Cox proportional hazards models.

²An asterisk superscript indicates that the estimate is significantly different than 1 (HR) or 0 (%Med)

Scientific Objective #6: Sensitivity Analysis: Were estimated treatment effects affected by impact of the COVID-19 pandemic on the study data?

Supplemental Figure 4. Hazard ratios and 95% confidence intervals for the treatment effects on the primary, secondary, and tertiary outcomes, based on the full data set and on a pre-COVID-19 data set restricted to data up to and including March 15, 2020.

A 3-panel figure, with a panel for the primary, secondary, and tertiary outcomes (from left to right). Pairwise hazard ratios (with 95% confidence intervals) for treatment effects on each outcome from an unadjusted Cox proportional hazards model will be estimated based on two data sets: (1) the full GRADE data set, and (2) a pre-COVID pandemic data set (i.e., including all data collected up to and including March 15, 2020). Since there are a total of 6 possible pairwise comparisons for each outcome, confidence intervals will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). A simple mocked-up version of a figure displaying these hazard ratios and 95% confidence intervals using simulated data is displayed below.

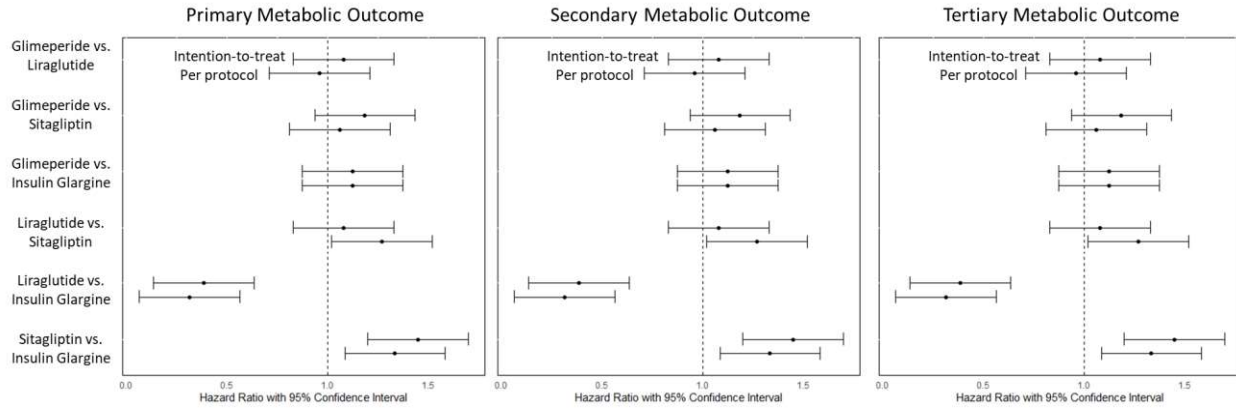


Scientific Objective #7: Sensitivity Analysis: Treatment effects among subset of GRADE data while on randomly assigned treatment (i.e., per-protocol analysis)

Supplemental Figure 5. Hazard ratios and 95% confidence intervals for the treatment effects on the primary, secondary, and tertiary outcomes, based on an intention-to-treat analysis and on a per-protocol analysis.

A 3-panel figure, with a panel for the primary, secondary, and tertiary outcomes (from left to right). Pairwise hazard ratios (with 95% confidence intervals) for treatment effects on each outcome from an unadjusted Cox proportional hazards model will be estimated based on two data sets: (1) the full analysis data set (intention-to-treat analysis), and (2) the per protocol data set (per protocol analysis). Since there are a total of 6 possible pairwise comparisons for each outcome, confidence intervals will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues”

section at the end of this document). A simple mocked-up version of a figure displaying these hazard ratios and 95% confidence intervals using simulated data is displayed below.



Scientific Objective #8: Sensitivity Analysis: Treatment effects if entire GRADE cohort had taken the assigned treatment according to study protocol during entire follow-up (inverse probability weighting analysis)

Supplemental Figure 6. Hazard ratios and 95% confidence intervals for the treatment effects on the primary, secondary, and tertiary outcomes, based on an intention-to-treat analysis and on an inverse probability weighting analysis to account for the impact of study treatment discontinuation.

A 3-panel figure, with a panel for the primary, secondary, and tertiary outcomes (from left to right). Pairwise hazard ratios (with 95% confidence intervals) for treatment effects on each outcome from an unadjusted Cox proportional hazards model will be estimated in the following two ways: (1) intention-to-treat analysis (i.e., using the full GRADE data set), and (2) inverse probability weighting (IPW) analysis.

The purpose of the IPW analysis is to estimate the treatment effects *if the entire GRADE cohort had taken the assigned treatment according to the study protocol during the entire study follow-up*. IP weights will be calculated based on fitted Cox proportional hazards models for the first treatment discontinuation greater than 4 weeks (28 days) in the full analysis data set. Then a weighted Cox proportional hazards model will be fit using the per protocol data set to estimate the treatment effects for each outcome using these IP weights and only including participants in the risk set prior to treatment discontinuation (i.e., considering any study visits occurring after treatment discontinuation as censored). The IP weight ($weight_{ij}$) for participant i at each discrete study time j will be calculated by

$$weight_{ij} = \prod_{k=0}^j \frac{P(D_{ik} = 0 | A_i = a_i, D_{is} = 0, s = 0, \dots, k-1)}{P(D_{ik} = 0 | A_i = a_i, C_{ik} = c_{ik}, D_{is} = 0, s = 0, \dots, k-1)}$$

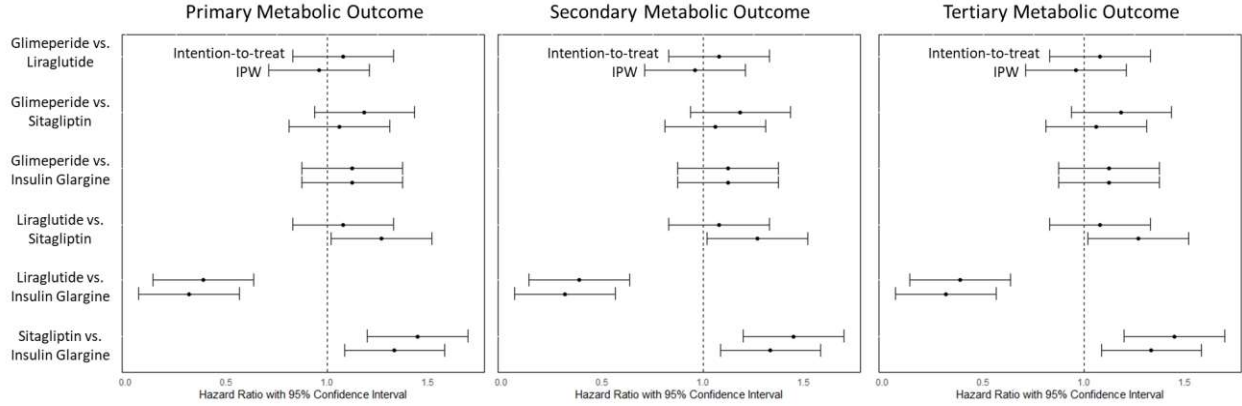
where D_{ik} indicates a binary indicator for treatment discontinuation (0 indicates no discontinuation) at study time k , A_i indicates the treatment group, and C_{ik} indicates a set of covariates related to discontinuation for study time k (potentially time-varying); a_i and c_{ik} indicate the observed values of A_i and C_{ik} respectively for participant i at study time k . The numerator and denominator can both be estimated using a Cox proportional hazards model given treatment group (for both the numerator and denominator) and the potentially time-varying covariates (for the denominator only). The *ipw* function in the *ipw* R package (van der Wal & Geskus, 2011) will be used to calculate the weights.

The following baseline covariates will be considered in the Cox proportional hazards model for treatment discontinuation: GRADE site, sex, race/ethnicity (combined), age, education, marital status, insurance status, employment status, and SF36 physical and social functioning score. In addition, the following time-varying covariates will be considered in the model for treatment discontinuation: metformin dose (<1000 mg/day, 1000-1999 mg/day, or 2000 mg/day), metformin type, lipids (total, LDL, HDL, and triglycerides), eGFR, Framingham score, metformin adherence (0% missed pills, 0-20% missed pills, or >20% missed pills), randomized medication dose, HbA1c, BMI, depression status, incidence of severe hypoglycemia, and incidence of a major SAE other than hypoglycemia. Interactions between treatment group and the following covariates should also be considered: lipids, eGFR, randomized medication dose, HbA1c, BMI, depression status, incidence of severe hypoglycemia, and incidence of a major SAE other than hypoglycemia. To reduce the complexity of the discontinuation model, a principal components analysis can be conducted to identify redundant effects in the model that can be removed. A condition index >30 will be used to identify collinear effects that should be addressed.

The treatment effect of interest will be estimated based on a weighted Cox proportional hazards model (applying the estimated IP weights) using the per protocol data set, where participants will be included in the risk set only up until the point that they first discontinue treatment for greater than 4 weeks (28 days). The IP weights of the participants who remain in the risk set will be repeatedly adjusted to account for discontinuations. Note that since the IP weights will be estimated rather than known, standard errors for the treatment effect of interest will be estimated using a robust standard error estimator.

Since there are a total of 6 possible pairwise comparisons for each outcome, confidence intervals will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document).

A simple mocked-up version of a figure displaying these hazard ratios and 95% confidence intervals using simulated data is displayed below.



STATISTICAL CONSIDERATIONS

Rationale for Non-Standard Statistical Methodology

Inverse probability weighting (IPW) sensitivity analysis

Inverse probability weighting (IPW) analyses will be conducted to estimate the treatment effect on the study outcomes *if the entire GRADE cohort had taken the assigned treatment according to the study protocol during the entire study follow-up* (i.e., Scientific Objective #8). These analyses will consist of (1) fitting a Cox proportional hazards model for the probability of treatment discontinuation conditional on covariates based on the full analysis set, (2) calculating IP weights based on the predicted probabilities of treatment discontinuation for each participant from that model, and (3) then fitting a weighted Cox proportional hazards model for the outcome using these IP weights, where participants will be included in the risk set only up until they discontinue treatment. Here we provide more detailed justification about how to calculate the IP weights.

Stabilized versions of IP weights will be used to reduce the variability of the weights. To adjust for time-varying treatment discontinuation, the stabilized IP weight ($weight_{ij}$) for participant i at each discrete study time j can be calculated by

$$weight_{ij} = \prod_{k=0}^j \frac{P(D_{ik} = 0 | A_i = a_i, \bar{D}_{ik} = \bar{d}_{ik})}{P(D_{ik} = 0 | A_i = a_i, C_{ik} = c_{ik}, \bar{D}_{ik} = \bar{d}_{ik})}$$

where D_{ik} indicates a binary indicator for treatment discontinuation (0 indicates no discontinuation) at study time k , \bar{D}_{ik} indicates the discontinuation history (i.e., all discontinuation indicators prior to study visit k), A_i indicates the treatment group, and C_{ik} indicates a set of covariates related to discontinuation for study time k (potentially time-varying); \bar{d}_{ik} , a_i , and c_{ik} indicate the observed values of \bar{D}_{ik} , A_i , and C_{ik} respectively for participant i at study time k (Robin et al, 2000; van der Wal & Geskus, 2011).

The factors in the denominator of $weight_{ij}$ correspond to the probability of continuing to use the assigned study treatment at study time k , given the treatment group, the current

covariates, and the prior discontinuation history. The purpose of the denominator is to adjust for differential treatment discontinuation by weighting the data subset with no treatment discontinuation to resemble the full GRADE cohort. In other words, study visits that are most similar (based on the covariate values) to visits with discontinued treatment (i.e., most similar to study visits that will be excluded from analysis) will have higher weights, and study visits that are least similar to visits with discontinued treatment will have lower weights.

The factors in the numerator of $weight_{ij}$ correspond to the probability of continuing to use the assigned study treatment at study time k , given the treatment group and the prior discontinuation history. The purpose of the numerator of the stabilized weight is to reduce the difference between the numerator and denominator of the weight, which reduces the variability of the weights (i.e., stabilizes the weights). This helps to avoid extremely large and influential weights.

Since we are considering the time to first treatment discontinuation only (i.e., instances where the participant does not return to the assigned study treatment at a later time), the stabilized IP weights can further be simplified as

$$weight_{ij} = \prod_{k=0}^j \frac{P(D_{ik} = 0 | A_i = a_i, D_{is} = 0, s = 0, \dots, k - 1)}{P(D_{ik} = 0 | A_i = a_i, C_{ik} = c_{ik}, D_{is} = 0, s = 0, \dots, k - 1)}$$

where the numerator and denominator can both be estimated using a Cox proportional hazards model given treatment group (for both the numerator and denominator) and the potentially time-varying covariates (for the denominator only) (van der Wal & Geskus, 2011). Note that since the IP weights will be estimated rather than known, standard errors for the treatment effect of interest will be estimated using a robust standard error estimator (Robins et al, 2000; van der Wal, 2011).

Other statistical issues

Significance level

A significance level of $\alpha=0.05$ will be used for all statistical tests, unless otherwise specified. Comparisons among the treatment groups will be adjusted for the number of tests conducted, 6 for pairwise comparisons and 4 for each group versus the average of the others. Unless stated otherwise, the adjusted p-values are obtained from application of the closed testing principle. In cases where the closed testing adjustment cannot be readily applied, then the Holm adjustment will be employed. Otherwise, p-values will be designated as “nominal” or “simple” p-values.

Intention-to-treat analyses

Unless otherwise specified, all available data for all randomized participants (i.e., the full analysis set) will be included in analyses, and data will be analyzed according to the randomly assigned treatment group, regardless of adherence to assigned treatment and/or compliance with the study protocol, according to intention-to-treat principles.

Definition of event times for glycemic outcomes (primary, secondary, and tertiary)

The event time will be defined based on the date of the triggering HbA1c value (not the date of the required confirmation value). An event time will be considered to be right censored at the final quarterly visit if the HbA1c is too low to trigger an outcome at the final visit (e.g., HbA1c <7% for primary outcome), and the confirmed outcome has not been reached at any point during follow-up. An event time will be considered to be right censored at the second-to-final quarterly visit if a triggering value of HbA1c was observed at the final visit (e.g., HbA1c \geq 7% for primary outcome), but HbA1c <7% at the second-to-final quarterly visit; this is because there is no confirmation value of HbA1c available following the triggering value at the final visit, and so it is unknown whether the event occurred at the final visit.

Checking the proportional hazards assumption for the Cox proportional hazards model

For analyses based on the Cox proportional hazards model, the assumption of proportional hazards will be tested using the test of Lin (Lin, 1991). If the test of proportional hazards is significant (i.e., hazards are assessed to be non-proportional), then the coefficients from the Cox model will be interpreted (approximately) as average log hazard ratios. Regardless of whether the proportional hazards assumption applies, inferences (standard errors, confidence intervals, and p-values) will be based on the robust information sandwich covariance estimates (Lin & Wei, 1989), and the robust model score test will be used to test for treatment group differences (Lachin, 2011).

Adjustments for multiple pairwise comparisons among the treatment groups

Since there are 4 treatment groups, there are 6 possible pairwise comparisons among the treatment groups (i.e., 6 elemental hypotheses of interest). A closed testing approach will be used to account for multiple pairwise comparisons among the treatment groups (Lachin et al, 2019). First, an omnibus T^2 -like test will be conducted to test for any differences among the 4 treatment groups; this is considered the order 3 hypothesis, which is the intersection of any 3 of the elemental hypotheses of pairwise differences. If that test is significant at the specified significance level α , then each of the order 2 sub-hypotheses (i.e., intersection hypotheses for 2 elemental hypotheses at a time) will be tested at significance level α . Each of the pairwise comparisons (i.e., order 1 hypotheses) can be tested at significance level α if all of the relevant higher-order hypotheses (i.e., order 3 and relevant order 2 hypotheses) are significant at significance level α . See the table below for an outline of the null hypotheses in the testing hierarchy that must be significant to allow for testing of each pairwise comparison (let $H_{0,1234}$ be the order 3 hypothesis that all 4 treatment groups are equal; $H_{0,ij,kl}$ be the order 2 hypothesis that treatment groups i and j are equal and treatment groups k and l are equal; $H_{0,ij}$ be the order 1 hypothesis that treatment groups i and j are equal).

Pairwise Comparison	Group 1 vs. 2	Group 1 vs. 3	Group 1 vs. 4	Group 2 vs. 3	Group 2 vs. 4	Group 3 vs. 4
Order 3 (4-group comparison)	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$
Order 2 (3-group comparison)	$H_{0,12,13}$	$H_{0,12,13}$	$H_{0,12,14}$	$H_{0,12,23}$	$H_{0,12,24}$	$H_{0,12,34}$
	$H_{0,12,14}$	$H_{0,13,14}$	$H_{0,13,14}$	$H_{0,13,23}$	$H_{0,13,24}$	$H_{0,13,34}$
	$H_{0,12,23}$	$H_{0,13,23}$	$H_{0,14,23}$	$H_{0,14,23}$	$H_{0,14,24}$	$H_{0,14,34}$
	$H_{0,12,24}$	$H_{0,13,24}$	$H_{0,14,24}$	$H_{0,23,24}$	$H_{0,23,24}$	$H_{0,23,34}$
	$H_{0,12,34}$	$H_{0,13,34}$	$H_{0,14,34}$	$H_{0,23,34}$	$H_{0,24,34}$	$H_{0,24,34}$
Order 1 (2-group comparison)	$H_{0,12}$	$H_{0,13}$	$H_{0,14}$	$H_{0,23}$	$H_{0,24}$	$H_{0,34}$

Comparing each treatment to all other treatments combined

There is interest in testing whether the effect of each treatment differs from the other 3 treatment groups combined. Let θ_k be the log(hazard ratio) comparing the hazard for treatment group $k = 1,2,3$ to the hazard for reference treatment group $k = 4$. For each treatment group, we would test the null hypothesis that the average of the estimated hazard ratios comparing each of the other treatments to the treatment of interest equals 1. In other words, we would test each of the following 4 null hypotheses (i.e., one hypothesis per treatment group):

$$H_{01}: \exp\{\theta_2 - \theta_1\} + \exp\{\theta_3 - \theta_1\} + \exp\{-\theta_1\} = 3$$

$$H_{02}: \exp\{\theta_1 - \theta_2\} + \exp\{\theta_3 - \theta_2\} + \exp\{-\theta_2\} = 3$$

$$H_{03}: \exp\{\theta_1 - \theta_3\} + \exp\{\theta_2 - \theta_3\} + \exp\{-\theta_3\} = 3$$

$$H_{04}: \exp\{\theta_1\} + \exp\{\theta_2\} + \exp\{\theta_3\} = 3$$

A closed testing approach will be used to account for multiple comparisons, according to the procedure described in (Lachin & Bebu, 2020) The closed testing hierarchy would start with the 3-df test of the joint hypothesis $\theta_1 = \theta_2 = \theta_3 = 0$. The next stage of the closed testing hierarchy would be to test the intersections of the elementary hypotheses listed above (e.g., $H_{01} \cap H_{02}$). The last stage would be to test the elementary hypotheses listed above. For example, the elementary hypothesis H_{01} would be rejected at significance level α if H_{01} , $H_{01} \cap H_{02}$, $H_{01} \cap H_{03}$, $H_{01} \cap H_{04}$, and the joint hypothesis $\theta_1 = \theta_2 = \theta_3 = 0$ are all significant at significance level α .

Adjustments for multiple comparisons for subgroup analyses

One of the objectives of this paper is to assess treatment group differences within baseline subgroups (e.g., tertiles of BMI). There are 6 possible pairwise comparisons among the treatment groups within each subgroup. A closed testing approach will also be used to account for multiple comparisons for testing treatment group differences within subgroups (Lachin et al, 2019). Here, we describe the general closed testing approach for the case with

all 4 treatment groups and 3 subgroups (e.g., tertiles of BMI), where θ_{jk} is the measure of treatment difference between treatment $k = 1,2,3$ and the reference treatment $k = 4$ within subgroup $j = a, b, c$. First, an overall test of the null hypothesis of homogeneity of treatment effects across all subgroups would be tested:

$$H_{0,abc}: \begin{aligned} \theta_{a1} &= \theta_{b1} = \theta_{c1} \\ \theta_{a2} &= \theta_{b2} = \theta_{c2} \\ \theta_{a3} &= \theta_{b3} = \theta_{c3} \end{aligned}$$

If this test is significant at the specified significance level ($\alpha = 0.05$), then tests of null hypotheses of homogeneity of treatment effects between pairs of subgroups would be tested:

$$H_{0,ab}: \theta_{a1} = \theta_{b1}, \theta_{a2} = \theta_{b2}, \theta_{a3} = \theta_{b3}$$

$$H_{0,ac}: \theta_{a1} = \theta_{c1}, \theta_{a2} = \theta_{c2}, \theta_{a3} = \theta_{c3}$$

$$H_{0,bc}: \theta_{b1} = \theta_{c1}, \theta_{b2} = \theta_{c2}, \theta_{b3} = \theta_{c3}$$

Then if any two of these tests were significant at the specified significance level ($\alpha = 0.05$), then within the intersection subgroup, tests of pairwise treatment comparisons can proceed in a similar manner as described in the previous section (related to adjustment of multiple pairwise comparisons among treatment groups). For example, if the tests of $H_{0,ab}$ and $H_{0,ac}$ were both significant, then testing of pairwise treatment comparisons can proceed within subgroup a .

Calculation of confidence intervals adjusted for multiple comparisons based on the closed testing framework

For analyses with multiple comparisons (e.g., pairwise treatment comparisons, comparisons of each treatment group vs. all others combined, subgroup analyses), confidence intervals for effect estimates will be calculated based on a method that controls the family-wise type 1 error for multiple comparisons.

APPENDIX A: Dataset Request

Table of Variables

This table defines the variables to be used in the analysis. The table has columns for the measure, the corresponding variable name in dataset, units, study visits at which the measure was collected, and notes for important details about the measure (e.g. standard study categories, definition if derived from other variables, etc.).

Measure	Variable	Units	Assessment Visits	Notes
Treatment	masked.trt		Baseline	

Glycemic Outcomes

Primary outcome	primaryEv, primaryYrs		Quarterly
Secondary outcome	secondaryEv, secondaryYrs		Quarterly
Tertiary outcome	tertiaryEv, tertiaryYrs		Quarterly

Other Clinical Variables

HbA1c	hba1c	%	Baseline, Quarterly	Categorized as tertiles for subgroup analyses
Fasting glucose	glu0	mg/dL	Baseline, 1-year Annual, 3-year Annual, 5-year Annual	
Weight	Weight	kg	Baseline, Quarterly	

Baseline Subgroup Variables

Race/ethnicity	race, Hispanic		Baseline	Categories: non-Hispanic white, non-Hispanic black, Hispanic white, other
Sex	Female		Baseline	Categories: male, female
Age	Age	years	Baseline	Categories: < 45 years, 45 - 59 years, 60+ years
Diabetes duration	diabDur.s	years	Screening	Categorized as tertiles
Body mass index (BMI)	Bmi	kg/m ²	Baseline, Quarterly	Categorized as tertiles

Study Compliance Variables

Attended close-out study visit	closeoutVisit			Close-out	
Visit adherence	visitAdherence	%	Quarterly	100% * (number of study visits attended)/(expected number of study visits according to study protocol)	
Duration of follow-up	fupTime	years	Baseline, Quarterly	Date of last study visit - randomization date	
Permanent discontinuation of metformin	discEventMet		Quarterly		
Permanent discontinuation of study treatment regimen	discEvent, discTime		Quarterly		
Temporary discontinuation of study treatment regimen	discEventTemp		Quarterly		
Off-study use of glucose-lowering medication	anyHiGluRx.long, surRx.long, dpp4Rx.long, gpl1Rx.long, insulinRx.long, sglT2Rx.long, otherHiGluRx.long		Quarterly	Overall, and specifically for off-study medications in the following classes: Sulfonylurea, DPP 4-inhibitor, GLP-1 RA, Insulin, SGLT-2 inhibitor	

Side Effects/Adverse Events

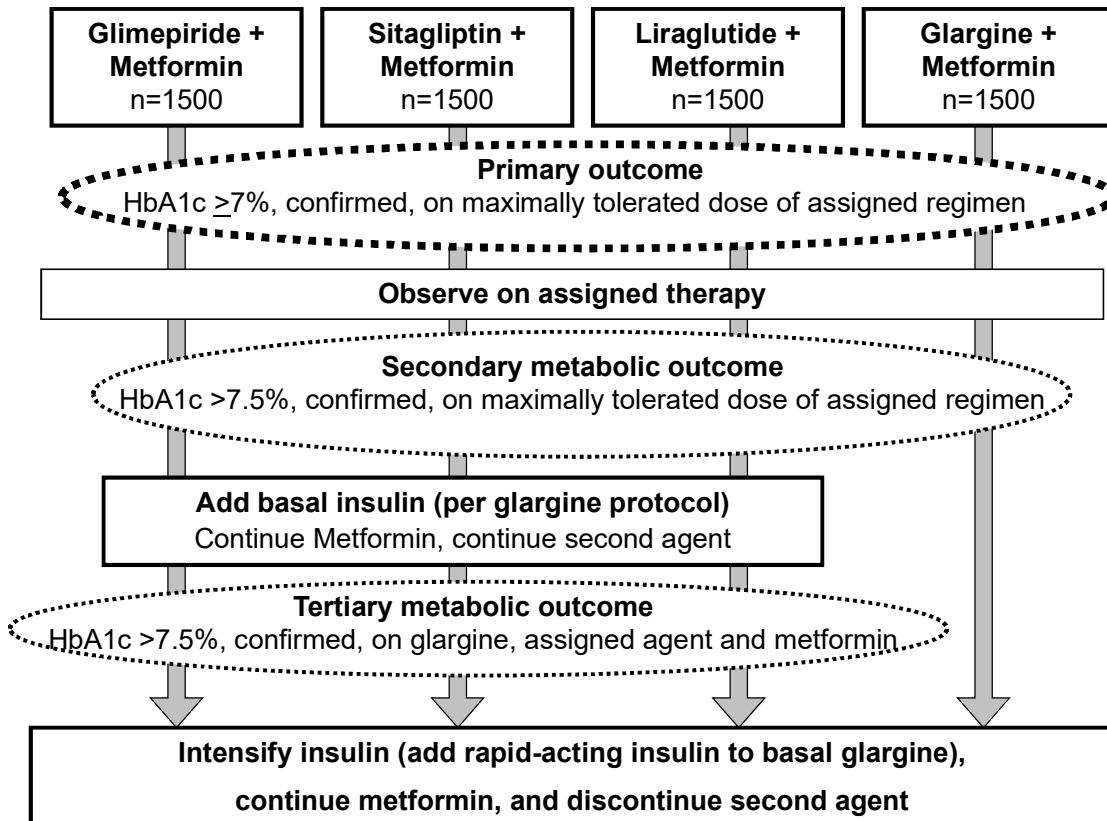
Mortality	deathEvent, deathNEvents, deathAtRisk
Any adverse event (targeted event or event resulting in	anyAEEEvent, anyAENEvents, anyAEAtRisk

hospitalization ≥ 24 hrs)		
Serious adverse event	SAEEvent, SAENEvents, SAEAtRisk	
Severe hypoglycemia	sevHypoEvent, sevHypoNEvents, sevHypoAtRisk	
Weight gain ≥ 10% higher than at randomization	wtPct10Event, wtPct10NEvents, wtPct10AtRisk	
Gastrointestinal symptoms	gastrohEvent, gastroNEvents, gastroAtRisk	Includes nausea, vomiting, diarrhea, stomach pain/bloating
Lactic acidosis	LAEvent, LANEvents, LAAAtRisk	Quarterly
Pancreatitis	PancreatitisEvent, PancreatitisNEvents, PancreatitisAtRisk	Quarterly
Acute metabolic decompensation	AMDEvent, AMDNEvents, AMDAtRisk	Includes diabetic ketoacidosis, HHS
Gallstone disease	gallstoneEvent, gallstoneNEvents, gallstoneAtRisk	Includes cholecystitis, cholelithiasis
Thyroid cancer	cancerThyroidEvent, cancerThyroidNEvents, cancerThyroidAtRisk, cancerMedullaryEvent, cancerMedullaryNEvents , cancerMedullaryAtRisk	All and medullary thyroid cancers
Pancreatic cancer	cancerPancreaticEvent, cancerPancreaticNEvent s, cancerPancreaticAtRisk	
Other cancer	cancerOtherEvent, cancerOtherNEvents, cancerOtherAtRisk	

Appendix B: Manuscript Figures Not Requiring Statistical Analysis

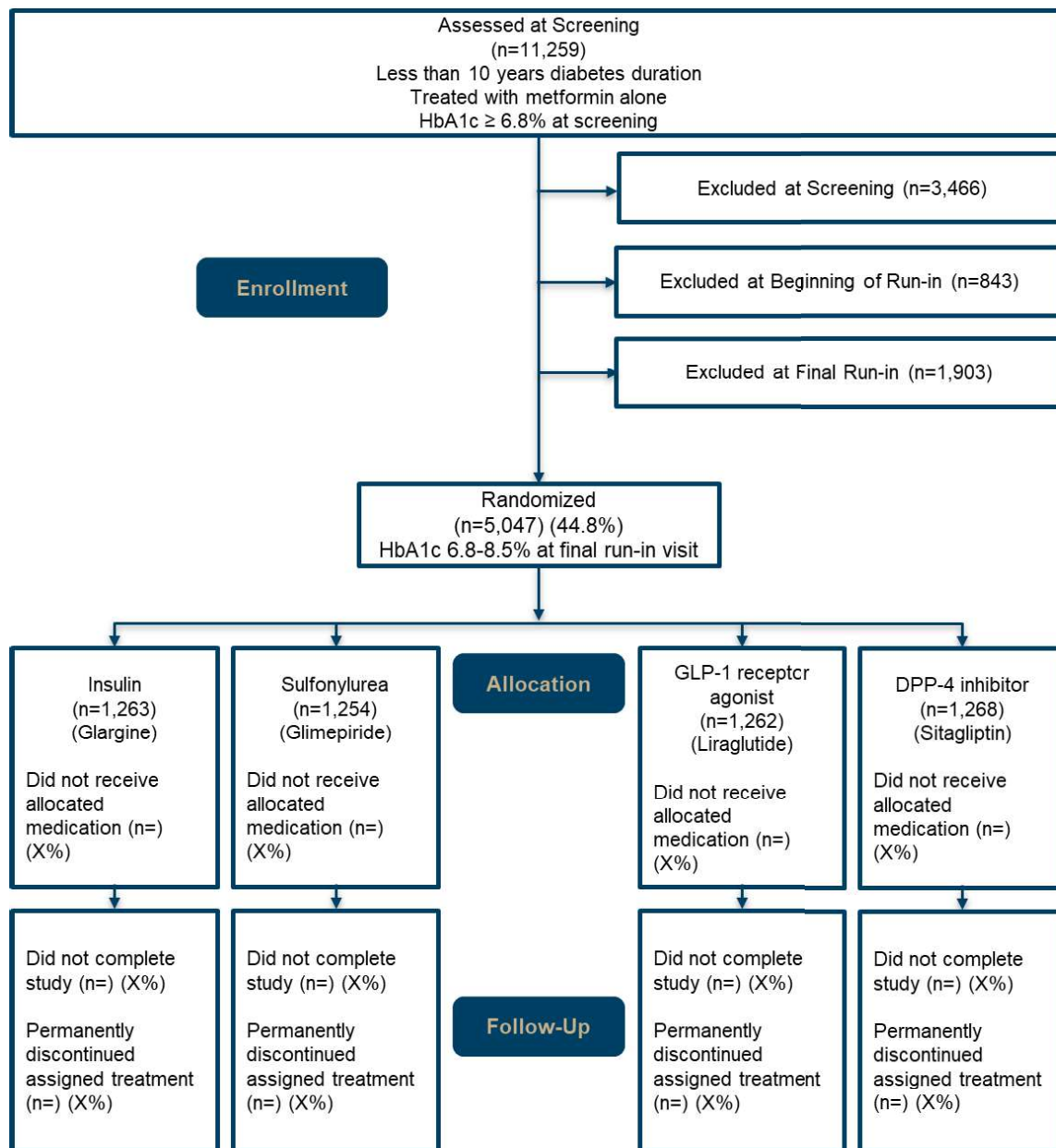
Supplemental Figure 1. Metabolic outcomes and subsequent therapy.

Will add in actual group sizes and numbers in each group that reach specific outcomes.



Supplemental Figure 2. Consolidated Standards of Reporting Trials (CONSORT) diagram

Note: Reasons for not receiving any dose of the allocated medication included the following: reason A (I: X%, G: X%, L: X%, S: X%), reason B (I: X%, G: X%, L: X%, S: X%),... Reasons for not completing the study included the following: death (I: X%, G: X%, L: X%, S: X%), withdrawal from study (I: X%, G: X%, L: X%, S: X%), loss to follow-up (I: X%, G: X%, L: X%, S: X%). Reasons for discontinuing the assigned treatment include the following: side effect or adverse event (I: X%, G: X%, L: X%, S: X%),...



REFERENCES

- 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.
- 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.

- Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; 11:561-570.
- Lachin JM, Bebu I. Closed testing of each group versus the others combined in a multiple group analysis. *Clinical Trials* 2020; 17:77-86.
- Lachin JM, Bebu I, Larsen MD, Younes N. Closed testing using surrogate hypotheses with restricted alternatives. *PLoS ONE* 2019; 14:1-18.
- Lachin, JM. *Biostatistical Methods: The Assessment of Relative Risks*. 2nd Edition. John Wiley and Sons; New York, 2011.
- 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* 1989; 84:1074-78.
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11:550-560.
- Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* 2014; 15:222-233.
- van der Wal WM, Geskus RB. ipw: An R package for inverse probability weighting. *Journal of Statistical Software* 2011; 43:1-23.

Amended Final Statistical Analysis Plan

Long term differences in metabolic status among four initial treatments added to metformin in early type 2 diabetes (OP1).

Table of Contents

GENERAL INFORMATION	2
APPROVALS.....	2
REVISION HISTORY.....	3
ABBREVIATIONS AND ACRONYMS.....	3
STUDY OBJECTIVES.....	3
Background and justification.....	3
Scientific objectives/questions	4
STATISTICAL METHODS AND DATASETS	4
Analysis Data Set Inclusion Criteria.....	4
Primary Variables to be Assessed.....	5
Statistical Analyses	7
Scientific Objective #1: Patient characteristics, retention, protocol completion, adherence by treatment group	7
Scientific Objective #2: Treatment effect on glycemic outcomes	10
Scientific Objective #3: Subgroup Analyses	14
Scientific Objective #4: Severe adverse events/targeted adverse events/side effects by treatment group.....	15
Scientific Objective #6: Sensitivity Analysis: Were estimated treatment effects affected by impact of the COVID-19 pandemic on the study data?	18
Scientific Objective #7: Sensitivity Analysis: Treatment effects among subset of GRADE data while on randomly assigned treatment (i.e., per-protocol analysis)	19
STATISTICAL CONSIDERATIONS.....	20
Other statistical issues	20
Significance level	20
Intention-to-treat analyses.....	20

Definition of event times for glycemic outcomes (primary, secondary, and tertiary) ..	20
Checking the proportional hazards assumption for the Cox proportional hazards model.....	20
Adjustments for multiple pairwise comparisons among the treatment groups.....	21
Comparing each treatment to all other treatments combined	21
Adjustments for multiple comparisons for subgroup analyses	22
Calculation of confidence intervals adjusted for multiple comparisons based on the closed testing framework.....	23
APPENDIX A: Dataset Request.....	23
Table of Variables.....	23
Appendix B: Manuscript Figures Not Requiring Statistical Analysis	26
REFERENCES	28

GENERAL INFORMATION

This document is an amended final statistical analysis plan for the above referenced manuscript. This document was prepared and reviewed by John M. Lachin, Naji Younes, Heidi Krause-Steinrauf and Nicole Butera.

Manuscript Title	Long term differences in metabolic status among four initial treatments added to metformin in early type 2 diabetes
GRADE paper number	OP1
Analysis Category	End of Study
Writing Group Chairs	David Nathan, John Lachin
Writing Group Members	David M. Nathan, John M. Lachin, Ashok Balasubramanyam, Henry B. Burch, John B. Buse, Nicole M. Butera, Robert M. Cohen, Jill P. Crandall, Steven E. Kahn, Heidi Krause-Steinrauf, Mary E. Larkin, Neda Rasouli, Margaret Tiktin, Deborah J. Wexler, Naji Younes
Target Journal	NEJM
Lead Statisticians	Naji Younes, Heidi Krause-Steinrauf, Nicole Butera

APPROVALS

No signatures are provided.

REVISION HISTORY

Version No.	Implemented by	Date	Reason
1	Nicole Butera	5/1/2021	Initial Version Approved
2	Nicole Butera	1/19/2022	Final SAP Implemented in Manuscript

ABBREVIATIONS AND ACRONYMS

Abbreviation	Meaning
HbA1c	Glycated hemoglobin
T2DM	Type 2 diabetes

STUDY OBJECTIVES

Background and justification

Type 2 diabetes (T2DM) affects more than 30 million persons in the United States, with an incidence of 1.5 million new cases per year, and more than 400 million persons world-wide. The major human and economic costs associated with T2DM are related primarily to the development of long-term diabetes-specific complications, including retinopathy, nephropathy, and neuropathy, and a 2-5 fold increased risk of non-specific cardiovascular disease (CVD). These long-term complications have been shown to be ameliorated in part by interventions that reduce chronic glycemia, as measured by glycated hemoglobin levels (HbA1c), and a target range of less than 7% (53 mmol/mol) has been established by consensus for most patients with T2DM. The estimated annual cost of diabetes in the US in 2017 was approximately \$327 billion dollars per year with an increasing fraction attributed to the cost of glucose-lowering medications.

Virtually all recommendations for the management of type 2 diabetes have included metformin as the first medication to be used. Unfortunately, choosing the second medication from the ever expanding list of glucose-lowering medications to add to metformin when monotherapy fails to achieve or maintain goal glycemia is problematic owing to the dearth of any long-term head-to-head comparator studies. The purpose of the Glycemia Reduction Approaches in Type 2 Diabetes: A Comparative Effectiveness (GRADE) Study was to examine the relative effectiveness of the four most commonly used glucose-lowering medications added to metformin to maintain goal glycemia. In this paper, we report the GRADE major glycemetic outcomes. The accompanying paper reports the vascular outcomes and CVD risk factors associated with the four randomly assigned interventions.

Scientific objectives/questions

1. Summarize patient characteristics, retention, protocol completion, and adherence across the four treatment groups.
2. Do the primary, secondary, and/or tertiary glyceic outcomes differ by treatment group?
3. Do treatment effects on the primary, secondary, and/or tertiary glyceic outcomes vary by the following pre-specified baseline subgroups: race/ethnicity, gender, age, diabetes duration, BMI, HbA1c?
4. Do severe adverse events/targeted adverse events/side effects differ by treatment group?
5. Are treatment effects on the primary, secondary, and/or tertiary glyceic outcomes mediated by other factors?
6. Sensitivity Analysis 1: Were the estimated treatment effects on the primary, secondary, and/or tertiary glyceic outcomes affected by the impact of the COVID-19 pandemic on the study data?
7. Sensitivity Analysis 2: What were the treatment effects on the primary, secondary, and/or tertiary glyceic outcomes among the subset of the GRADE data while on the randomly assigned treatment (i.e., per-protocol analysis)?
8. Sensitivity Analysis 3: What would the treatment effects on the primary, secondary, and/or tertiary glyceic outcomes have been if the entire GRADE cohort had taken the assigned treatment according to study protocol during the entire follow-up period?

STATISTICAL METHODS AND DATASETS

Analysis Data Set Inclusion Criteria

Full analysis set: All available follow-up data from all randomized participants (n=5047).

For scientific objective #7, will use the *per-protocol* data set:

- The subset of participants who meet both of the following criteria:
 - Took at least one dose of the assigned therapy
 - Completed at least one outcome assessment visit
- The subset of participant data that meets the following criteria:
 - Data up to the end of study for patients who do not permanently discontinue the assigned drug regimen during the study and/or initiate the use of non-study diabetes drug(s).

- Data prior to the first permanent discontinuation of the assigned drug regimen, for patients who so discontinued. A subject is considered to have discontinued from the study regimen (i.e., the assigned medications according to the study protocol) if the subject stops taking at least one of the study medications called for under the regimen (e.g., a subject who fails to start glargine after reaching the secondary outcome is considered to have discontinued the study regimen). In particular, note that data following discontinuation of the randomly assigned medication after reaching the tertiary outcome would be excluded from the per-protocol dataset.
- Data prior to initiation of use of non-study diabetes drug(s).
- Data up to the time of withdrawal from the study.

Primary Variables to be Assessed

Treatment assignment: Glimepiride (Sulfonylurea), Liraglutide (GLP-1 RA), Sitagliptin (DPP 4-inhibitor), Glargine (Insulin)

HbA1c during follow-up

Glycemic outcomes (*definitions from study protocol*):

- **Primary glycemic outcome:** Time to an initial HbA1c $\geq 7\%$, subsequently confirmed at the next quarterly visit. If the initially observed HbA1c is $> 9\%$, then the confirmation value will be performed within 3 to 6 weeks. If the initial HbA1c and confirmation value 3 to 6 weeks later are both $> 9\%$, the primary and secondary outcomes will have been reached. If the initial HbA1c is $> 9\%$ and the confirmation value 3 to 6 weeks later is $\leq 9\%$, the participant will resume his/her usual schedule of quarterly HbA1c monitoring. If the HbA1c at the next quarterly visit is $\geq 7\%$, then the primary outcome will have been reached. The primary outcome can only be reached after a minimum of 6 months of therapy, unless the HbA1c at 3 months is $> 9\%$ and is higher for the confirmation HbA1c 3-6 weeks later, in which case the primary and secondary outcomes will have been met at 3 months.
- **Secondary glycemic outcome:** Time to an HbA1c $> 7.5\%$ after having reached the primary outcome, subsequently confirmed at the next quarterly visit. The primary and secondary outcomes may be reached simultaneously if the initial value and the confirmation are both $> 7.5\%$.
- **Tertiary glycemic outcome:** Time to an HbA1c $> 7.5\%$ after having confirmed the secondary outcome (at which point the participant should have started basal insulin based on study protocol), subsequently confirmed at the next quarterly visit. Note that following the intention-to-treat framework, the main analyses for this paper will define the tertiary outcome irrespective of whether participants actually start basal insulin following a secondary outcome according to study protocol (*based on decision made during Writing Group call on 11/24/2020*).

Study compliance variables:

- **Visit adherence:** $100\% * (\text{number of study visits attended}) / (\text{expected number of study visits according to study protocol})$. The denominator (i.e., expected number of study visits according to study protocol) includes the number of visits based on the amount of time elapsed between randomization and the expected close-out visit date (based on the date of randomization) for those who survived to the end of the study or date of death.
- **Duration of follow-up:** Date of last study contact minus randomization date
- **Discontinuation of metformin:** The participant reports permanently stopping metformin.
- **Off-study use of glucose-lowering medication and/or discontinuation of study treatment regimen:** Use of non-study, off-protocol glucose-lowering medication at a study visit and/or stopping at least one of the study medications called for based on the study protocol. At least one of the study medications has been stopped permanently.
- **Discontinuation of study treatment regimen:** Stopping at least one of the study medications called for based on the study protocol. The assigned study treatment regimen has been stopped permanently.
- **Off-study use of glucose-lowering medication:** Overall, and specifically for off-study medications in the following classes: sulfonylurea, DPP 4-inhibitor, GLP-1 RA, insulin, SGLT-2 inhibitor, thiazolidinedione, other

Adverse events and side effects:

- Mortality
- Any adverse event (targeted event or event resulting in hospitalization overnight or ≥ 24 hours)
- Serious adverse event
- Hospitalization overnight or ≥ 24 hours
- Severe or major hypoglycemia
 - Severe hypoglycemia requires 3rd party assistance
 - Major hypoglycemia is a severe episode that results in loss of consciousness and/or seizure
 - Severe hypoglycemia that results in injury to the participant or others (e.g. motor vehicle accident in which the participant was the driver)
- Weight gain $\geq 10\%$ higher than at randomization

- Gastrointestinal symptoms (nausea, vomiting, diarrhea, stomach pain/bloating)
- Lactic acidosis
- Pancreatitis
- Acute metabolic decompensation (diabetic ketoacidosis, HHS)
- Gallstone disease (cholelithiasis, cholecystitis)
- Cancer (thyroid, pancreatic, other)

NOTE: a table listing of all variables is provided in the Appendix hereto.

Statistical Analyses

Scientific Objective #1: Patient characteristics, retention, protocol completion, adherence by treatment group

Supplemental Table 1. Selected baseline characteristics related to the glycemic outcomes.

Continuous row variables: Mean (SD) for row variable overall and stratified by treatment group, unless otherwise specified.

Binary/categorical row variables: n (%) for row variable overall and stratified by treatment group.

	All	Insulin Glargine	Glimepiride	Liraglutide	Sitagliptin
n (%)	xxx	xxx	xxx	xxx	xxx
Age					
<45 years	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
45-59 years	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
≥60 years	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Women (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Race					
Am Ind/Alaska Native	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Asian	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)

Hawaiian/Pacific Isl	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Black or African-American	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
White	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Other/multiple	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Unknown/not reported	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Ethnicity	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Duration of diabetes (years), Mean (SD)	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Duration of diabetes (years), Median (IQR)	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]
Baseline metformin dose					
1000	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
1500	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
2000	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
BMI (kg/m ²)	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
HbA1c	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x

Table 1. Retention, protocol completion and adherence comparing the treatment groups during the entire GRADE study period.

Discontinuation of assigned treatment off-protocol refers to permanently stopping at least one of the study medications called for based on the study protocol (e.g., a participant who fails to start glargine after reaching the secondary outcome is considered to have discontinued the assigned treatment, whereas stopping the randomized medication due to reaching the tertiary outcome is not considered to have discontinued).

Continuous row variables: Mean (SD) for row variable overall and stratified by treatment group, unless otherwise specified.

Binary/categorical row variables: n (%) for row variable overall and stratified by treatment group.

	All	Insulin Glargine	Glimepiride	Liraglutide	Sitagliptin
--	-----	---------------------	-------------	-------------	-------------

N	xxx	xxx	xxx	xxx	xxx
Attended close-out study visit (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Visit adherence (%) ¹	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Mean duration of follow-up (years) ²	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Median duration of follow-up (years) ²	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]	x.x [x.x, x.x]
Number (%) of participants who discontinued metformin	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Number (%) that used non-study, off-protocol glucose-lowering medications and/or discontinued assigned study treatment regimen off-protocol ³	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Within 1st year post-randomization	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
1 - 2 years post-randomization	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
2+ years post-randomization	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Time (years) on assigned study treatment regimen per protocol ³	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
% of study time on assigned study treatment regimen per protocol ^{3,5}	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x	x.x±x.x
Number (%) that discontinued assigned study treatment regimen off-protocol ⁴	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Use of non-study, off-protocol glucose-lowering medications (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Sulfonylurea (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
DPP 4-inhibitor (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)

GLP-1 RA (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Insulin (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
SGLT-2 inhibitor (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Thiazolidinedione (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)
Other (%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)	xxx (x.x%)

¹ Visit adherence = 100% * (number of study visits attended) / (expected number of study visits according to the study protocol), calculated for each individual

² Duration of follow-up = date of last study contact – date of randomization

³ Only includes treatment discontinuation that was not consistent with the study protocol. Specifically, this does not include discontinuation of the randomized medication due to reaching the tertiary outcome, as required by study protocol.

⁴ Participants were considered to have discontinued the assigned treatment regimen if the study medication(s) were discontinued for a minimum of 4 weeks.

⁵ Percent of time from randomization to the expected close-out visit date (based on the date of randomization) for those who survived to the end of the study or date of death.

Scientific Objective #2: Treatment effect on glycemc outcomes

Figure 1. Cumulative incidence of (a) primary, (b) secondary and (c) tertiary glycemc outcomes by treatment group. Mean (d) HbA1c, (e) fasting plasma glucose levels and (f) weight over study time.

The 3 panels in the top row display the cumulative incidence for the primary, secondary, and tertiary outcomes (from left to right) over time. Each panel includes 4 lines, one for the cumulative incidence within each treatment group. The cumulative incidence by treatment group will be estimated using a Kaplan-Meier estimator. The total number at risk at each year will be provided below each panel. The time axis will represent the time since GRADE randomization. The maximum value for the time axis will be selected as the last time when the total number at risk is ≥ 200 for the primary outcome. The y-axis limits will be selected to be the same for the panels for the primary, secondary, and tertiary outcomes, per journal requirements.

An additional panel displays the mean values of HbA1c (%) over time. HbA1c was collected at each quarterly visit, and so means will be displayed for every 3 months. 95% confidence bands for the longitudinal means will be graphed, based on a simple repeated measures model for the longitudinal means without covariates. This panel includes 4 lines, one for longitudinal means with each treatment group. For consistency, the same time axis will be

used for this panel as for the top panels displaying cumulative incidence of the primary, secondary, and tertiary glycemic outcomes. The number of participants at each year will be provided below each panel.

A simple mocked-up version of this figure using simulated data is displayed below.

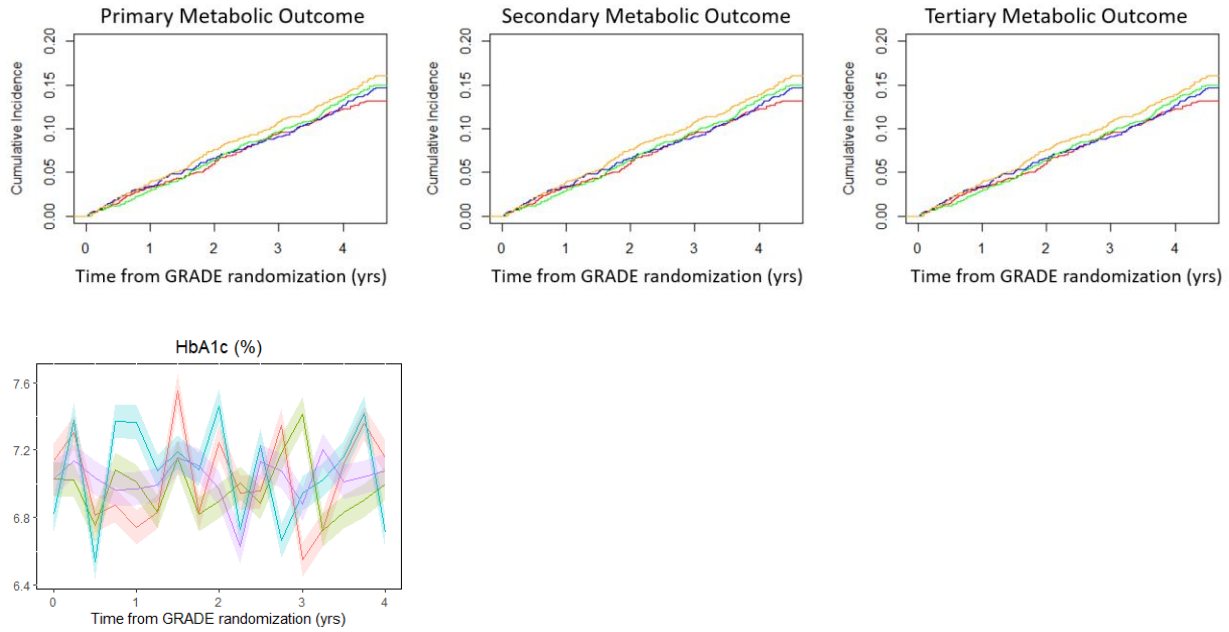
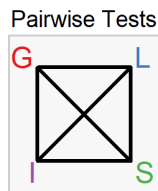


Table 2. Numbers of subjects reaching primary, secondary and tertiary glycemic outcomes by treatment group, with crude rates, pairwise hazard ratios, and hazard ratios compared to all other treatments combined.



For this table, the following statistics will be calculated for the primary, secondary, and tertiary glycemic outcomes, both overall and stratified by treatment group:

- The number of events and percent of the GRADE cohort with the outcome.
- Crude rate per 100 person-years (SE). This will be calculated as $100 \times (\text{observed number of events}) / (\text{total time at risk})$, where the total time at risk is the sum of the time since randomization to the event (or to the censoring time for those without an event) across participants.
- Pairwise hazard ratios (SE). A Cox proportional hazards model will be fit for the outcome with treatment group as a predictor. For the purposes of this Cox model, the event times and censoring times will be calculated as time since randomization to the event or censoring respectively. Hazard ratios and standard errors for each pairwise comparison of the treatment groups will be estimated from the Cox model. All Wald-type tests, standard errors and confidence intervals will be estimated using the robust Lin-Wei (1989) information sandwich estimator to ensure valid inferences even if the

proportional hazards assumption does not apply. A joint test for differences in the hazards among any of the treatment groups will be conducted. If that joint test is significant, then pairwise log-rank tests will be conducted to test for all pairwise differences. There are a total of 6 possible pairwise comparisons among the 4 treatment groups, and therefore these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). If the joint test for differences among any of the treatment groups is significant, then the results from the pairwise testing will be visualized using the following graphic, where each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly; dotted lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and solid lines indicate $p \leq 0.001$.



- Hazard ratio compared to all other treatments combined (SE) (Lachin and Bebu, 2020). A Cox proportional hazards model will be fit for the outcome with treatment group as a predictor. For the purposes of this Cox model, the event times and censoring times will be calculated as time since randomization to the event or censoring respectively. For treatment a , the hazard ratio compared to all other treatments combined will be estimated as the average of the estimated hazard ratios comparing each of the other treatments to treatment a . Since there are 4 treatment groups, there would be a total of 4 tests comparing each treatment to all others combined, and therefore these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document).
- RMST (SE). The restricted mean survival time (RMST) up to $\tau = 4$ years will be estimated, including standard errors.
- According to the study protocol, participants should add glargine insulin to their treatment regimen after reaching the secondary outcome. Therefore, this table will also report the number and percent of participants who actually start treatment with glargine insulin following the secondary outcome. The percent will be calculated as $100\% * (\text{number of participants who started glargine insulin}) / (\text{number of participants /who reached the secondary outcome})$.

	Total (N=5047)	Glimepiride (G) (N=XXXX)	Liraglutide (L) (N=XXXX)	Sitagliptin (S) (N=XXXX)	Insulin Glargine (I) (N=XXXX)	Pairwise Treatment Comparisons ¹
Primary Metabolic Outcome						
n (%)						
Crude rate per 100 person- years (SE)						
Pairwise hazard ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 
Hazard ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
RMST (SE)						
Secondary Metabolic Outcome						
n (%)						
Crude rate per 100 person- years (SE)						
Pairwise hazard ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	-- -- --		Pairwise Tests 
Hazard ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
RMST (SE)						
Number (%) starting on insulin after reaching	--				--	

secondary outcome						
Tertiary Metabolic Outcome						
n (%)						
Crude rate per 100 person-years (SE)	--					
Pairwise hazard ratios (SE)						
Glimepiride (SE) Liraglutide (SE) Sitagliptin (SE)	-- -- --	-- -- --	-- -- --	--		
Hazard ratio compared to all other treatments combined (SE)	--					G: p= L: p= S: p= I: p=
RMST (SE)						

* $p \leq 0.05$ (from test that hazard ratio equals 1)

** $p \leq 0.01$ (from test that hazard ratio equals 1)

*** $p \leq 0.001$ (from test that hazard ratio equals 1)

¹ Boxes in this column graphically display the results of testing pairwise treatment effects. Each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly. Solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

Scientific Objective #3: Subgroup Analyses

Subgroup analyses of the treatment effects for the primary, secondary, and tertiary glycemic outcomes within subgroups based on the following baseline variables: race, ethnicity, sex, age, diabetes duration, body mass index, and HbA1c. The subgroups for each baseline variable are defined as the following:

- Race: White, black, and other (the “other” category includes all race categories other than white or black, and includes participants who specified other or unknown race)
- Ethnicity: Hispanic/Latino, non-Hispanic/Latino
- Sex: Male, female
- Age: <45 years, 45-59 years, 60+ years
- Diabetes duration, body mass index, HbA1c: Sample tertiles

Figure 2. Analyses for primary outcome within protocol pre-specified baseline subgroups

Graphs of the cumulative incidence of the primary outcome will be presented stratified by the subgroup variable. Since there are a total of 6 possible pairwise comparisons, these tests will be adjusted for multiple comparisons using a Holm testing procedure (see details in the “Other statistical issues” section at the end of this document).

Similar figures will be produced for the secondary outcome and the tertiary outcome also.

Scientific Objective #4: Severe adverse events/targeted adverse events/side effects by treatment group

Table 3. Overall incidence and rate of adverse events and side effects with comparison of treatment groups



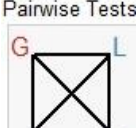



This table displays the number and percent of participants who experienced each type of adverse event/side effect, and the crude rate (SE) of the adverse event/side effect, stratified by treatment group. The crude rates per 100 person-years will be calculated as $100 * (\text{observed number of events}) / (\text{total time at risk})$.







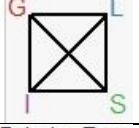
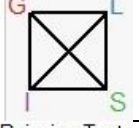

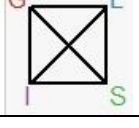
Tests of all pairwise treatment comparisons for each adverse event/side effect will be assessed, based on a Poisson regression of the number of events with the log of exposure time as an offset and treatment as the only covariate. Since there are a total of 6 possible pairwise comparisons for each adverse event/side effect, these tests will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). The results from the pairwise testing will be visualized using a graphic, where each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly; solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.





The following adverse events/side effects will be included in this table:

- Mortality
- Any adverse event (targeted event or event resulting in hospitalization ≥ 24 hours)
- Serious adverse event
- Hospitalization overnight or ≥ 24 hours
- Severe or major hypoglycemia
- Weight gain $\geq 10\%$ higher than at randomization
- Gastrointestinal symptoms (nausea, vomiting, diarrhea, stomach pain/bloating)
- Lactic acidosis
- Pancreatitis

- Acute metabolic decompensation (diabetic ketoacidosis, HHS)
- Gallstone disease (cholecystitis and cholelithiasis)
- Thyroid cancer (all, medullary)
- Pancreatic cancer
- Other cancer

Adverse Event	Glimepiride (G) (N=XXXX)		Liraglutide (L) (N=XXXX)		Sitagliptin (S) (N=XXXX)		Insulin Glargine (I) (N=XXXX)		Pairwise Treatment Comparisons ₁
	n (%)	Rate (SE) ²	n (%)	Rate (SE) ²	n (%)	Rate (SE) ²	n (%)	Rate (SE) ²	
Mortality ³									Pairwise Tests 
Any adverse event (targeted event or event resulting in hospitalization ≥ 24 hours)									Pairwise Tests 
Serious adverse event									Pairwise Tests 
Hospitalization overnight or ≥ 24 hours									
Severe hypoglycemia ^{3,4}									Pairwise Tests 
Major hypoglycemia ^{3,5}									Pairwise Tests 
Weight gain ≥10% higher than at randomization									Pairwise Tests 

Gastrointestinal symptoms									Pairwise Tests 
Nausea									Pairwise Tests 
Vomiting									Pairwise Tests 
Diarrhea									Pairwise Tests 
Stomach pain / bloating									Pairwise Tests 
Lactic acidosis ³									Pairwise Tests 
Pancreatitis ³									Pairwise Tests 
Acute metabolic decompensation									Pairwise Tests 
Diabetic ketoacidosis									Pairwise Tests 
Hyperosmolar Hyperglycemic Syndrome									Pairwise Tests 

Gallstone disease (cholecystitis and cholelithiasis)									Pairwise Tests 
Thyroid cancer ³ All Medullary									Pairwise Tests 
Pancreatic cancer ³									Pairwise Tests 
Other cancer ³									Pairwise Tests 

¹ Boxes in this column graphically display the results of testing pairwise treatment effects. Each corner of the box represents one of the four treatments (G = Glimepiride, L = Liraglutide, S = Sitagliptin, I=Insulin Glargine), and lines connect the treatments that differ significantly. Solid lines indicate $p \leq 0.05$, dashed lines indicate $p \leq 0.01$, and dotted lines indicate $p \leq 0.001$.

² Event rate per 100 person-years. HHS- hyperglycemic hyperosmolar syndrome.

³ Adjudicated events

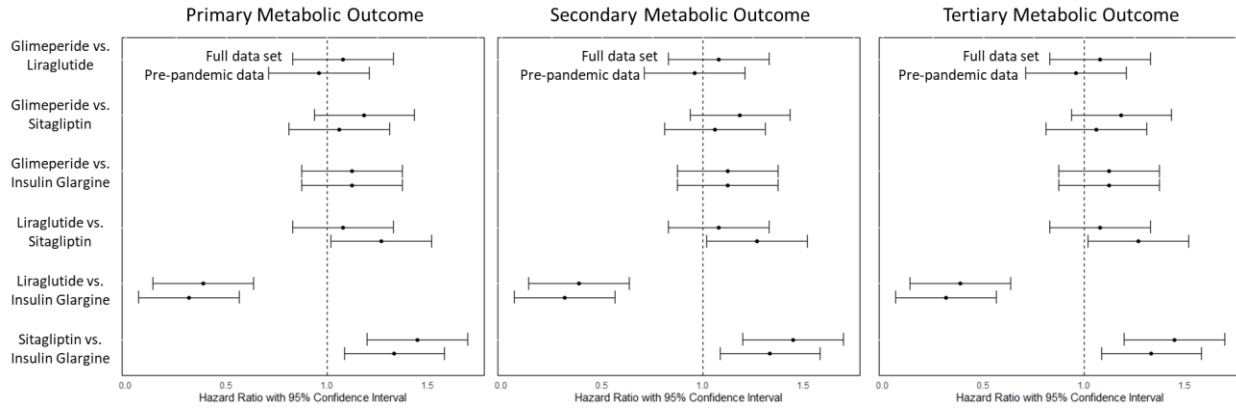
⁴ Defined as an episode requiring third-party assistance to treat

⁵ Defined as an episode resulting in coma/seizure

Scientific Objective #6: Sensitivity Analysis: Were estimated treatment effects affected by impact of the COVID-19 pandemic on the study data?

Supplemental Figure 4. Hazard ratios and 95% confidence intervals for the treatment effects on the primary, secondary, and tertiary outcomes, based on the full data set and on a pre-COVID-19 data set restricted to data up to and including March 15, 2020.

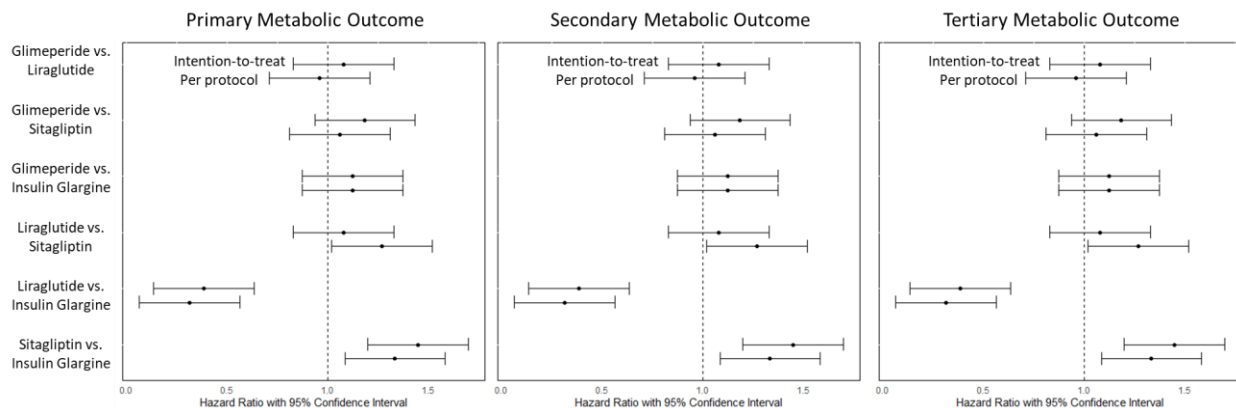
A 3-panel figure, with a panel for the primary, secondary, and tertiary outcomes (from left to right). Pairwise hazard ratios (with 95% confidence intervals) for treatment effects on each outcome from an unadjusted Cox proportional hazards model will be estimated based on two data sets: (1) the full GRADE data set, and (2) a pre-COVID pandemic data set (i.e., including all data collected up to and including March 15, 2020). Since there are a total of 6 possible pairwise comparisons for each outcome, confidence intervals will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). A simple mocked-up version of a figure displaying these hazard ratios and 95% confidence intervals using simulated data is displayed below.



Scientific Objective #7: Sensitivity Analysis: Treatment effects among subset of GRADE data while on randomly assigned treatment (i.e., per-protocol analysis)

Supplemental Figure 5. Hazard ratios and 95% confidence intervals for the treatment effects on the primary, secondary, and tertiary outcomes, based on an intention-to-treat analysis and on a per-protocol analysis.

A 3-panel figure, with a panel for the primary, secondary, and tertiary outcomes (from left to right). Pairwise hazard ratios (with 95% confidence intervals) for treatment effects on each outcome from an unadjusted Cox proportional hazards model will be estimated based on two data sets: (1) the full analysis data set (intention-to-treat analysis), and (2) the per protocol data set (per protocol analysis). Since there are a total of 6 possible pairwise comparisons for each outcome, confidence intervals will be adjusted for multiple comparisons using a closed testing procedure (see details in the “Other statistical issues” section at the end of this document). A simple mocked-up version of a figure displaying these hazard ratios and 95% confidence intervals using simulated data is displayed below.



STATISTICAL CONSIDERATIONS

Other statistical issues

Significance level

A significance level of $\alpha=0.05$ will be used for all statistical tests, unless otherwise specified. Comparisons among the treatment groups will be adjusted for the number of tests conducted, 6 for pairwise comparisons and 4 for each group versus the average of the others. Unless stated otherwise, the adjusted p-values are obtained from application of the closed testing principle. In cases where the closed testing adjustment cannot be readily applied, then the Holm adjustment will be employed. Otherwise, p-values will be designated as “nominal” or “simple” p-values.

Intention-to-treat analyses

Unless otherwise specified, all available data for all randomized participants (i.e., the full analysis set) will be included in analyses, and data will be analyzed according to the randomly assigned treatment group, regardless of adherence to assigned treatment and/or compliance with the study protocol, according to intention-to-treat principles.

Definition of event times for glycemic outcomes (primary, secondary, and tertiary)

The event time will be defined based on the date of the triggering HbA1c value (not the date of the required confirmation value). An event time will be considered to be right censored at the final quarterly visit if the HbA1c is too low to trigger an outcome at the final visit (e.g., HbA1c <7% for primary outcome), and the confirmed outcome has not been reached at any point during follow-up. An event time will be considered to be right censored at the second-to-final quarterly visit if a triggering value of HbA1c was observed at the final visit (e.g., HbA1c \geq 7% for primary outcome), but HbA1c <7% at the second-to-final quarterly visit; this is because there is no confirmation value of HbA1c available following the triggering value at the final visit, and so it is unknown whether the event occurred at the final visit.

Checking the proportional hazards assumption for the Cox proportional hazards model

For analyses based on the Cox proportional hazards model, the assumption of proportional hazards will be tested using the test of Lin (Lin, 1991). If the test of proportional hazards is significant (i.e., hazards are assessed to be non-proportional), then the coefficients from the Cox model will be interpreted (approximately) as average log hazard ratios. Regardless of whether the proportional hazards assumption applies, inferences (standard errors, confidence intervals, and p-values) will be based on the robust information sandwich covariance estimates (Lin & Wei, 1989), and the robust model score test will be used to test for treatment group differences (Lachin, 2011).

Adjustments for multiple pairwise comparisons among the treatment groups

Since there are 4 treatment groups, there are 6 possible pairwise comparisons among the treatment groups (i.e., 6 elemental hypotheses of interest). A closed testing approach will be used to account for multiple pairwise comparisons among the treatment groups (Lachin et al, 2019). First, an omnibus T^2 -like test will be conducted to test for any differences among the 4 treatment groups; this is considered the order 3 hypothesis, which is the intersection of any 3 of the elemental hypotheses of pairwise differences. If that test is significant at the specified significance level α , then each of the order 2 sub-hypotheses (i.e., intersection hypotheses for 2 elemental hypotheses at a time) will be tested at significance level α . Each of the pairwise comparisons (i.e., order 1 hypotheses) can be tested at significance level α if all of the relevant higher-order hypotheses (i.e., order 3 and relevant order 2 hypotheses) are significant at significance level α . See the table below for an outline of the null hypotheses in the testing hierarchy that must be significant to allow for testing of each pairwise comparison (let $H_{0,1234}$ be the order 3 hypothesis that all 4 treatment groups are equal; $H_{0,ij,kl}$ be the order 2 hypothesis that treatment groups i and j are equal and treatment groups k and l are equal; $H_{0,ij}$ be the order 1 hypothesis that treatment groups i and j are equal).

Pairwise Comparison	Group 1 vs. 2	Group 1 vs. 3	Group 1 vs. 4	Group 2 vs. 3	Group 2 vs. 4	Group 3 vs. 4
Order 3 (4-group comparison)	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$	$H_{0,1234}$
Order 2 (3-group comparison)	$H_{0,12,13}$	$H_{0,12,13}$	$H_{0,12,14}$	$H_{0,12,23}$	$H_{0,12,24}$	$H_{0,12,34}$
	$H_{0,12,14}$	$H_{0,13,14}$	$H_{0,13,14}$	$H_{0,13,23}$	$H_{0,13,24}$	$H_{0,13,34}$
	$H_{0,12,23}$	$H_{0,13,23}$	$H_{0,14,23}$	$H_{0,14,23}$	$H_{0,14,24}$	$H_{0,14,34}$
	$H_{0,12,24}$	$H_{0,13,24}$	$H_{0,14,24}$	$H_{0,23,24}$	$H_{0,23,24}$	$H_{0,23,34}$
	$H_{0,12,34}$	$H_{0,13,34}$	$H_{0,14,34}$	$H_{0,23,34}$	$H_{0,24,34}$	$H_{0,24,34}$
Order 1 (2-group comparison)	$H_{0,12}$	$H_{0,13}$	$H_{0,14}$	$H_{0,23}$	$H_{0,24}$	$H_{0,34}$

Comparing each treatment to all other treatments combined

There is interest in testing whether the effect of each treatment differs from the other 3 treatment groups combined. Let θ_k be the log(hazard ratio) comparing the hazard for treatment group $k = 1,2,3$ to the hazard for reference treatment group $k = 4$. For each treatment group, we would test the null hypothesis that the average of the estimated hazard ratios comparing each of the other treatments to the treatment of interest equals 1. In other words, we would test each of the following 4 null hypotheses (i.e., one hypothesis per treatment group):

$$H_{01}: \exp\{\theta_2 - \theta_1\} + \exp\{\theta_3 - \theta_1\} + \exp\{-\theta_1\} = 3$$

$$H_{02}: \exp\{\theta_1 - \theta_2\} + \exp\{\theta_3 - \theta_2\} + \exp\{-\theta_2\} = 3$$

$$H_{03}: \exp\{\theta_1 - \theta_3\} + \exp\{\theta_2 - \theta_3\} + \exp\{-\theta_3\} = 3$$

$$H_{04}: \exp\{\theta_1\} + \exp\{\theta_2\} + \exp\{\theta_3\} = 3$$

A closed testing approach will be used to account for multiple comparisons, according to the procedure described in (Lachin & Bebu, 2020) The closed testing hierarchy would start with the 3-df test of the joint hypothesis $\theta_1 = \theta_2 = \theta_3 = 0$. The next stage of the closed testing hierarchy would be to test the intersections of the elementary hypotheses listed above (e.g., $H_{01} \cap H_{02}$). The last stage would be to test the elementary hypotheses listed above. For example, the elementary hypothesis H_{01} would be rejected at significance level α if $H_{01}, H_{01} \cap H_{02}, H_{01} \cap H_{03}, H_{01} \cap H_{04}$, and the joint hypothesis $\theta_1 = \theta_2 = \theta_3 = 0$ are all significant at significance level α .

Adjustments for multiple comparisons for subgroup analyses

One of the objectives of this paper is to assess treatment group differences within baseline subgroups (e.g., tertiles of BMI). There are 6 possible pairwise comparisons among the treatment groups within each subgroup. A Holm testing approach will be used to account for multiple comparisons for testing whether treatment group comparisons differ across subgroups. Here, we describe the general approach for the case with all 4 treatment groups and 3 subgroups (e.g., tertiles of BMI), where θ_{jk} is the measure of treatment difference between treatment $k = 1,2,3$ and the reference treatment $k = 4$ within subgroup $j = a, b, c$. First, an overall test of the null hypothesis of homogeneity of treatment effects across all subgroups would be tested:

$$H_{0,abc}: \begin{aligned} \theta_{a1} &= \theta_{b1} = \theta_{c1} \\ \theta_{a2} &= \theta_{b2} = \theta_{c2} \\ \theta_{a3} &= \theta_{b3} = \theta_{c3} \end{aligned}$$

If this test is significant at the specified significance level ($\alpha = 0.05$), then tests of null hypotheses of homogeneity of the 6 pairwise differential treatment effects is tested:

$$H_{0,12}: \theta_{a12} = \theta_{b12} = \theta_{c12} \text{ [Glargine (k=1) v Glimepiride (k=2)]}$$

$$H_{0,13}: \theta_{a13} = \theta_{b13} = \theta_{c13} \text{ [Glargine (k=1) v Liraglutide (k=3)]}$$

$$H_{0,14}: \theta_{a14} = \theta_{b14} = \theta_{c14} \text{ [Glargine (k=1) v Sitagliptin (k=4)]}$$

$$H_{0,23}: \theta_{a23} = \theta_{b23} = \theta_{c23} \text{ [Glimepiride (k=2) v Liraglutide (k=3)]}$$

$$H_{0,24}: \theta_{a24} = \theta_{b24} = \theta_{c24} \text{ [Glimepiride (k=2) v Sitagliptin (k=4)]}$$

$$H_{0,34}: \theta_{a34} = \theta_{b34} = \theta_{c34} \text{ [Liraglutide (k=3) v Sitagliptin (k=4)]}$$

The p-values from these 6 tests of homogeneity are holm adjusted for 6-tests. The final adjusted p-value for each pairwise test of homogeneity is then taken as the maximum of these holm adjusted p-values and the p-value from the overall test of homogeneity described above.

Then the 6 pairwise tests within each subgroup are conducted using a Holm adjustment for the number of subgroups, e.g. 18 tests with 3 subgroups.

Calculation of confidence intervals adjusted for multiple comparisons based on the closed testing framework

For analyses with multiple comparisons (e.g., pairwise treatment comparisons, comparisons of each treatment group vs. all others combined, subgroup analyses), confidence intervals for effect estimates will be calculated based on a method that controls the family-wise type 1 error for multiple comparisons.

APPENDIX A: Dataset Request

Table of Variables

This table defines the variables to be used in the analysis. The table has columns for the measure, the corresponding variable name in dataset, units, study visits at which the measure was collected, and notes for important details about the measure (e.g. standard study categories, definition if derived from other variables, etc.).

Measure	Variable	Units	Assessment Visits	Notes
Treatment	masked.trt		Baseline	
Glycemic Outcomes				
Primary outcome	primaryEv, primaryYrs		Quarterly	
Secondary outcome	secondaryEv, secondaryYrs		Quarterly	
Tertiary outcome	tertiaryEv, tertiaryYrs		Quarterly	
Other Clinical Variables				
HbA1c	hba1c	%	Baseline, Quarterly	Categorized as tertiles for subgroup analyses
Fasting glucose	glu0	mg/dL	Baseline, 1-year	

Weight	Weight	kg	Annual, 3-year Annual, 5-year Annual Baseline, Quarterly	
Baseline Subgroup Variables				
Race	race		Baseline	Categories: white, black, other
Ethnicity	Hispanic		Baseline	Categories: Hispanic/Latino, non-Hispanic/Latino
Sex	Female		Baseline	Categories: male, female
Age	Age	years	Baseline	Categories: < 45 years, 45 - 59 years, 60+ years
Diabetes duration	diabDur.s	years	Screening	Categorized as tertiles
Body mass index (BMI)	Bmi	kg/m ²	Baseline, Quarterly	Categorized as tertiles
Study Compliance Variables				
Attended close-out study visit	closeoutVisit		Close-out	
Visit adherence	visitAdherence	%	Quarterly	100% * (number of study visits attended)/(expected number of study visits according to study protocol)
Duration of follow-up	fupTime	years	Baseline, Quarterly	Date of last study visit - randomization date
Discontinuation of metformin	discEventMet		Quarterly	

Off-study use of glucose-lowering medication and/or discontinuation of study treatment regimen	discEvent, discTime, anyHiGluRx.long	Quarterly	
Discontinuation of study treatment regimen	discEvent	Quarterly	
Off-study use of glucose-lowering medication	anyHiGluRx.long, surRx.long, dpp4Rx.long, gpl1Rx.long, insulinRx.long, sglT2Rx.long, otherHiGluRx.long	Quarterly	Overall, and specifically for off-study medications in the following classes: Sulfonylurea, DPP 4-inhibitor, GLP-1 RA, Insulin, SGLT-2 inhibitor

Side Effects/Adverse Events

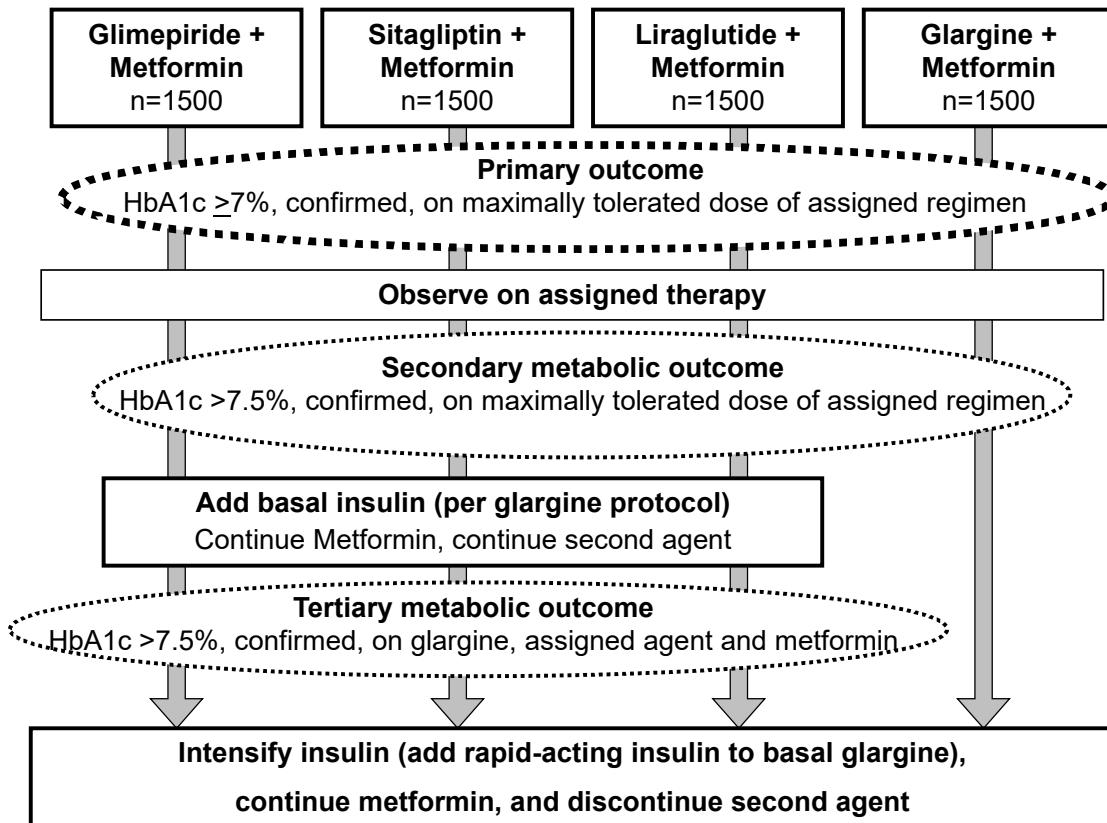
Mortality	deathEvent, deathNEvents, deathAtRisk
Any adverse event (targeted event or event resulting in hospitalization ≥ 24 hrs)	anyAEEEvent, anyAENEvents, anyAEAtRisk
Serious adverse event	SAEEEvent, SAENEvents, SAEAtRisk
Severe hypoglycemia	sevHypoEvent, sevHypoNEvents, sevHypoAtRisk
Weight gain ≥ 10% higher than at randomization	wtPct10Event, wtPct10NEvents, wtPct10AtRisk

Gastrointestinal symptoms	gastrohEvent, gastroNEvents, gastroAtRisk		Includes nausea, vomiting, diarrhea, stomach pain/bloating
Lactic acidosis	LAEvent, LANEvents, LAAtRisk	Quarterly	
Pancreatitis	PancreatitisEvent, PancreatitisNEvents, PancreatitisAtRisk	Quarterly	
Acute metabolic decompensation	AMDEvent, AMDNEvents, AMDAtRisk		Includes diabetic ketoacidosis, HHS
Gallstone disease	gallstoneEvent, gallstoneNEvents, gallstoneAtRisk		Includes cholecystitis, cholelithiasis
Thyroid cancer	cancerThyroidEvent, cancerThyroidNEvents, cancerThyroidAtRisk, cancerMedullaryEvent, cancerMedullaryNEvents, cancerMedullaryAtRisk		All and medullary thyroid cancers
Pancreatic cancer	cancerPancreaticEvent, cancerPancreaticNEvents, cancerPancreaticAtRisk		
Other cancer	cancerOtherEvent, cancerOtherNEvents, cancerOtherAtRisk		

Appendix B: Manuscript Figures Not Requiring Statistical Analysis

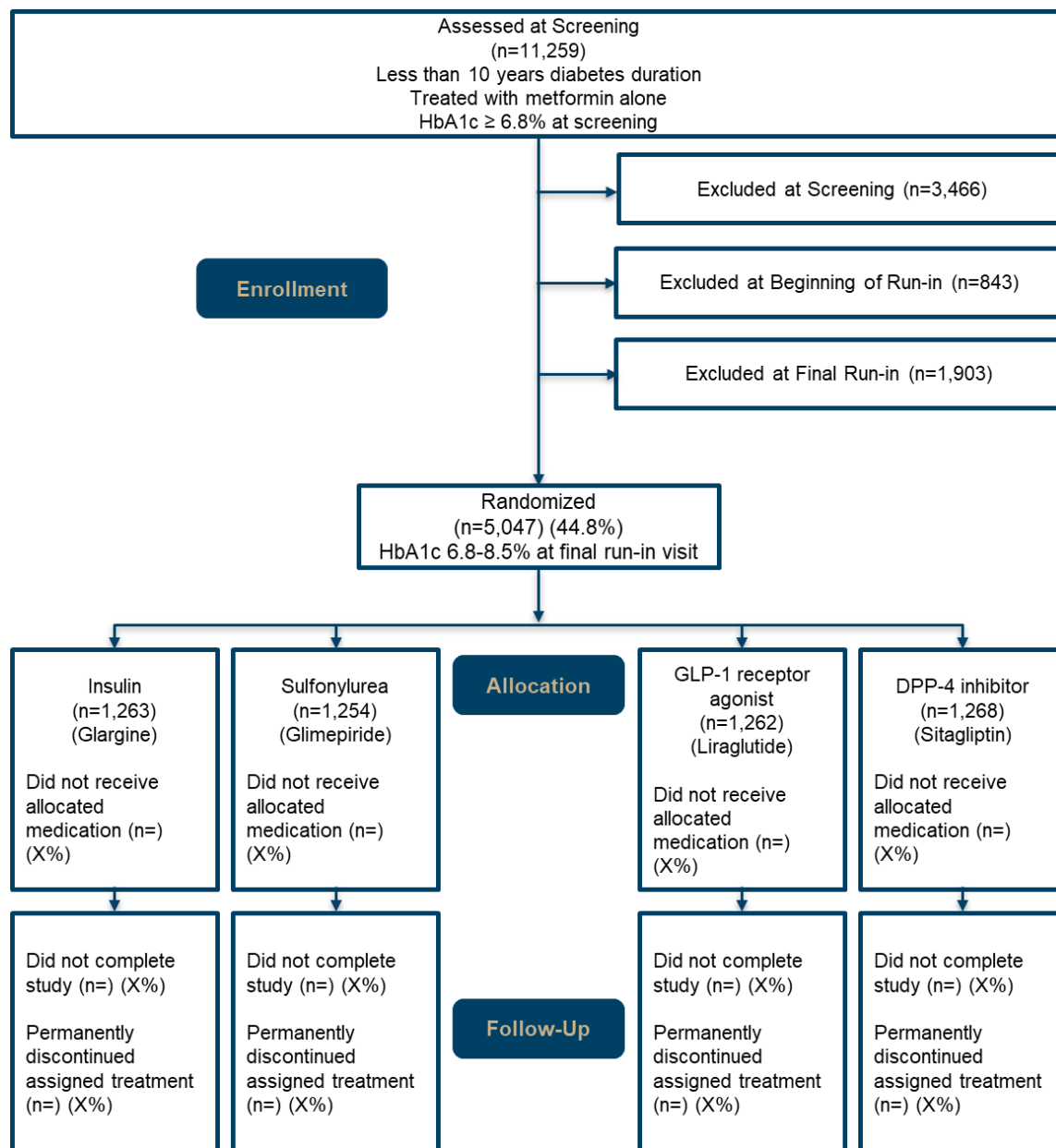
Supplemental Figure 1. Metabolic outcomes and subsequent therapy.

Will add in actual group sizes and numbers in each group that reach specific outcomes.



Supplemental Figure 2. Consolidated Standards of Reporting Trials (CONSORT) diagram

Note: Reasons for not receiving any dose of the allocated medication included the following: reason A (I: X%, G: X%, L: X%, S: X%), reason B (I: X%, G: X%, L: X%, S: X%),... Reasons for not completing the study included the following: death (I: X%, G: X%, L: X%, S: X%), withdrawal from study (I: X%, G: X%, L: X%, S: X%), loss to follow-up (I: X%, G: X%, L: X%, S: X%). Reasons for discontinuing the assigned treatment include the following: side effect or adverse event (I: X%, G: X%, L: X%, S: X%),...



REFERENCES

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.

Lachin JM, Bebu I. Closed testing of each group versus the others combined in a multiple group analysis. *Clinical Trials* 2020; 17:77-86.

Lachin JM, Bebu I, Larsen MD, Younes N. Closed testing using surrogate hypotheses with restricted alternatives. *PLoS ONE* 2019; 14:1-18.

Lachin, JM. *Biostatistical Methods: The Assessment of Relative Risks*. 2nd Edition. John Wiley and Sons; New York, 2011.

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* 1989; 84:1074-78.

Supplement to the Statistical Analysis Plan

Long term differences in metabolic status among four initial treatments added to metformin in early type 2 diabetes (OP1)

The above referenced paper is the first of two parallel manuscripts to be submitted for publication of the principal results of the Glycemia Reduction Approaches in Type 2 Diabetes: A Comparative Effectiveness (GRADE) Study. The Statistical Analysis Plan was completed on May 1, 2021 and signed by the lead investigators during May 5-11 of 2021. The last patient visit occurred around May 1. Data acquisition was expected to continue beyond that point, principally from the central laboratory and reading units.

Around May 1, a preliminary data lock was conducted to serve as the basis for analyses to be conducted as the basis for a 2 hour presentation of preliminary results during the virtual 2021 meeting of the American Diabetes Association on June 28, 2021. When conducting these analyses we decided to deviate in some respects from the specifications in the Statistical Analysis Plan. Herein each such deviation is described along with an explanation or justification.

1. *Original SAP: The per protocol analysis set was planned to exclude data prior to the first discontinuation of the assigned drug regimen for greater than 4 weeks (28 days).*

Deviation: The per protocol analysis set excluded data prior to permanent discontinuation of the assigned drug regimen.

Justification: It was decided that this definition for the per protocol analysis set is more clinically meaningful.

2. *Original SAP: Visit adherence (presented in Table 1, defined on p.6 of the original SAP) was defined as $100 * (\text{number of study visits attended}) / (\text{expected number of study visits according to study protocol})$. The expected number of study visits according to the study protocol was not explicitly defined in the SAP.*

Deviation: The denominator in this definition includes the number of visits based on the amount of time elapsed between randomization and the expected close-out visit date (based on the date of randomization) for those who survived to the end of the study or date of death.

Justification: This is a clarification of how the denominator of visit adherence was defined, since this definition was not specific in the original SAP.

3. *Original SAP: Table 1 was planned to present the following variables related to discontinuation of assigned study treatment regimen and/or use of non-study, off-protocol glucose-lowering medications:*
 - *Permanent discontinuation of assigned study treatment regimen off-protocol (i.e., not taking protocol-specified medication at all subsequent study visits).*

- *Temporary discontinuation of assigned study treatment regimen off-protocol (for greater than 4 weeks).*
- *Use of non-study, off-protocol glucose-lowering medication at a study visit.*

Deviation: Instead, the following variables related to discontinuation of assigned study treatment regimen and/or use of non-study, off-protocol glucose-lowering medications were included in Table 1:

- Use of non-study, off-protocol glucose-lowering medication at a study visit and/or permanent discontinuation of assigned study treatment regimen off-protocol.
- Permanent discontinuation of assigned study treatment regimen off-protocol.
- Use of non-study, off-protocol glucose-lowering medication at a study visit.

All variables related to discontinuation of the study treatment regimen were defined based on permanent treatment discontinuation.

Justification: To be more consistent with the definition of the pre-specified per protocol analyses. The number of participants who used non-study, off-protocol glucose-lowering medication at a study visit and/or discontinued the assigned study treatment regimen off-protocol corresponded directly to the number of participants who had at least a subset of their data excluded from the per protocol data set.

4. *Original SAP: Percent of study time on assigned study treatment regimen per protocol (defined in footnote 5 of Table 1 of SAP) was originally planned to be defined as percent of time from randomization to date of last study contact calculated for each individual.*

Deviation: The denominator in this percentage is now defined as the amount of time elapsed between randomization and the expected close-out visit date (based on the date of randomization) for those who survived to the end of the study or date of death.

Justification: To be more consistent with the definition of visit adherence (i.e., the denominators for visit adherence and percent of study time on assigned study treatment regimen per protocol are defined similarly).

5. *Original SAP: Figure 1 was originally planned to include graphs for the mean values of HbA1c, fasting glucose, and weight over time, by treatment group. It was also originally planned to present graphs for the kernel-smoothed distributions of HbA1c, fasting glucose, and weight by treatment group at 1 year and 3 years post-randomization. See p.11 – 12 in the SAP for details.*

Deviation: Graphs for the mean values of fasting glucose and weight were excluded from the manuscript. Graphs for the kernel-smoothed distributions of HbA1c, fasting glucose, and weight at 1 year and 3 years post-randomization were also excluded from the manuscript.

Justification: Space limitations/too much content for one paper.

6. *Original SAP: Table 2 was originally planned to include analyses of restricted mean survival times (RMST). Pairwise RMST ratios and RMST ratios compared to all other treatments combined were to be presented so as to quantify the treatment effects on the RMST of the primary, secondary, and tertiary outcomes.*

Deviation: RMST ratios among treatment groups (both pairwise RMST ratios and RMST ratios compared to all other treatments combined) were excluded from Table 2. Instead, the RMST for each treatment group was provided in the table.

Justification: The RMST results were used as descriptive analyses rather than for testing hypotheses, since hypothesis tests related to hazard ratios were also presented as the primary comparison among groups. Thus, it was decided not to present two different sets of hypothesis tests comparing treatment effects based on the same data.

7. Original SAP: It was originally planned to additionally present graphs of cumulative incidence and hazard ratio results for the secondary and tertiary outcome, where time was based on the time elapsed since the trigger HbA1c value for the previous outcome. See p.16 – 18 in the SAP for more details.

Deviation: These results (where time was based on time elapsed since the trigger HbA1c value for the previous outcome) were excluded from the manuscript.

Justification: Space limitations/too much content for one paper.

8. Original SAP: For the subgroup analyses, originally a closed testing procedure was planned for each subgroup variable to protect type 1 error due to multiple testing. The hypothesis tests of interest were specified as the six pairwise tests comparing the four treatment groups within each subgroup category (for example, the effect of glargine vs. glimepiride among white participants). Details of this closed testing procedure specified for subgroup analyses in the original SAP can be found on p.34 – 35 of the SAP.

Deviation: The testing procedure for subgroup analyses was implemented in the following way so as to protect the type 1 error probability due to multiple testing. First, an overall test of the null hypothesis of homogeneity of treatment effects across all categories of the subgroup variable was conducted. This overall test of homogeneity of treatment effects served as a gate-keeper test, where if this overall test of homogeneity was not significant for a given subgroup variable, then no further testing was done for that subgroup variable. If this overall test of homogeneity was significant, then a Holm testing procedure was used to protect type 1 error due to pairwise testing for that subgroup variable. The hypothesis tests of interest for each subgroup variable consisted of six pairwise tests for the four treatment groups, where each test assessed whether the pairwise treatment effect comparison differed significantly across subgroups (for example, whether the effect of glargine vs. glimepiride differed across race categories).

Then the 6 pairwise tests within each strata are conducted using a Holm adjustment for the number of strata, e.g. 18 tests with 3 strata.

Justification: This set of hypothesis tests addressed a more clinically meaningful research question than the set of hypothesis tests described in the original analysis plan. In addition, the closed testing procedure for this setting is not available in existing statistical software.

9. Original SAP: For subgroup analyses, it was planned to conduct analyses stratified by a combined race/ethnicity variable, where the categories would be the following: non-Hispanic white, non-

Hispanic black, Hispanic white, and other (see the description of the race/ethnicity subgroup variable on p.19 of the SAP).

Deviation: In the subgroup analyses presented in the paper, race (white, black, and other/multiple) and ethnicity (Hispanic/Latino vs. not Hispanic/Latino) were considered as separate subgroup variables.

Justification: Based on discussion with the writing group, it was decided that race and ethnicity are separate constructs, and so should be analyzed separately. This is also in keeping with the NIH-recommended racial and ethnic categories to be employed in descriptions of diversity (Notice Number: NOT-OD-15-089).

10. *Original SAP:* For presentation of the subgroup analysis results, it was planned to present forest plots of the crude rates of the primary outcome for each treatment group within each subgroup category (see p.19 – 20 of the SAP). It was also planned to present these subgroup analysis results in a table (see p.21 – 23 of the SAP).

Deviation: For the subgroup analysis results presented in the paper, graphs of the cumulative incidence were presented stratified by the subgroup variable, instead of the originally planned tables and figures.

Justification: This was a more clinically meaningful and parsimonious way to present the results.

11. *Original SAP:* Mediation analyses were planned to estimate the proportion of treatment effects on the glycemic outcomes that are explained by weight as a mediator (see p.27 in the SAP).

Deviation: These mediation analyses were not conducted for this paper.

Justification: Based on discussion with the writing group, it was decided (prior to conducting any mediation analyses) that mediation analyses would be too complex to add to this paper, given the amount of content already included in this paper, and that it would be better to address mediation analyses in a separate paper at a later time.

12. *Original SAP:* A sensitivity analysis was planned to use inverse probability weighting to estimate the treatment effects if the entire GRADE cohort had taken the assigned treatment according to the study protocol during the entire study follow-up (i.e., if no one had discontinued the assigned study treatment regimen off-protocol); see p.29 – 31 in the SAP.

Deviation: These sensitivity analyses using inverse probability weighting were not conducted for this paper.

Justification: Since the inverse probability weighting analyses would be complex and complicated to explain, and given the amount of content already included in this paper, it was decided that sensitivity analyses using inverse probability weighting would instead be conducted for a separate paper that will focus on sensitivity analyses for the main results in GRADE. Further, the pre-planned per-protocol sensitivity analyses are presented in this paper with a summary in the text and details in the Supplemental Material.