

Emulating the GRADE Trial Using Real-World Data

Yihong Deng, PhD¹; Eric C. Polley, PhD²; Joshua D. Wallach, PhD, MS³; Sanket S. Dhruva, MD, MHS^{4,5}; Jeph Herrin, PhD^{6,7}; Kenneth Quinto, MD, MPH⁹; Charu Gandotra, MD¹⁰; William Crown, PhD¹¹; Peter Noseworthy, MD¹²; Xiaoxi Yao, PhD, MPH^{1,12}; Timothy D. Lyon, MD¹³ Nilay D. Shah, PhD¹⁴; Joseph S. Ross, MD, MHS^{15,16}; Rozalina G. McCoy, MD, MS^{1,17}

- 1) Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN
- 2) Department of Public Health Sciences, University of Chicago, Chicago, IL
- 3) Department of Environmental Health Sciences, Yale School of Public Health, New Haven, CT
- 4) Section of Cardiology, San Francisco Veterans Affairs Health Care System, San Francisco, CA
- 5) Department of Medicine, UCSF School of Medicine, San Francisco, CA
- 6) Section of Cardiovascular Medicine, Yale School of Medicine, New Haven, CT
- 7) Flying Buttress Associates, Charlottesville, VA
- 8) Office of Medical Policy, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Springs, MD
- 9) Office of New Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Springs, MD
- 10) Florence Heller Graduate School, Brandeis University, Waltham, MA
- 11) Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN
- 12) Department of Urology, Mayo Clinic, Jacksonville, FL

13) Delta Air Lines, Atlanta, GA

14) Department of Internal Medicine, Yale School of Medicine, New Haven, CT

15) Department of Health Policy and Management, Yale School of Public Health, New Haven, CT

16) Division of Community Internal Medicine, Geriatrics, and Palliative Care, Department of Medicine, Mayo Clinic, Rochester, MN

Corresponding author: Rozalina G. McCoy, MD MS. Division of Community Internal Medicine, Geriatrics, and Palliative Care, Department of Medicine, Mayo Clinic. 200 First Street SW. Rochester, MN 55905. Phone (507) 284-5160; fax (507) 266-0036; e-mail mccoy.rozalina@mayo.edu.

Word count: Abstract 323, main text 4574, 3 tables, 1 figures, 40 references

ABSTRACT

OBJECTIVES. To emulate the Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE) trial using real-world data prior to its publication. GRADE is the first comparative effectiveness study to directly compare second-line glucose-lowering medications with respect to their ability to lower hemoglobin A_{1c} (HbA_{1c}).

DESIGN AND SETTING. In this observational cohort study, we applied GRADE trial criteria to claims and laboratory data from OptumLabs[®] Data Warehouse (OLDW), a U.S. nationwide claims database, between 1/25/2010 and 6/30/2019.

PARTICIPANTS. Adults with type 2 diabetes with hemoglobin A_{1c} (HbA_{1c}) 6.8-8.5% on metformin monotherapy, identified according to GRADE trial specifications.

INTERVENTIONS. Glimepiride, liraglutide, sitagliptin, and insulin glargine.

MAIN OUTCOME MEASURES. The primary outcome was time to HbA_{1c} $\geq 7.0\%$ and secondary outcomes were other metabolic, microvascular, macrovascular, and safety outcomes specified by GRADE. Propensity scores were estimated using the gradient boosting machine method and inverse propensity score weighting was used to emulate randomization of the treatment groups, which were then compared using Cox proportional hazards regression.

RESULTS. We identified 8252 patients (19.7% of adults starting the study drugs in OLDW) meeting GRADE eligibility criteria (glimepiride arm=4318, liraglutide arm=690, sitagliptin arm=2993, glargine arm=251). The glargine arm was excluded from analyses due to small sample size. Median times to HbA_{1c} $\geq 7.0\%$ were 442 (95% CI, 394-480) days for glimepiride, 764 (95% CI, 741-NA) days for liraglutide, and 427 (95% CI, 380-483) days for sitagliptin. Liraglutide was associated with lower risk of reaching HbA_{1c} $\geq 7.0\%$ compared to glimepiride (HR 0.57 [95% CI, 0.43-0.75]) and sitagliptin (HR 0.55 [95% CI, 0.41-0.73]). Results were

consistent for the secondary outcome of time to HbA_{1c} >7.5%. There were no significant differences among treatment groups for the remaining secondary outcomes.

CONCLUSIONS. In this emulation of the GRADE trial, liraglutide was significantly more effective at maintaining glycemic control than glimepiride or sitagliptin when added to metformin monotherapy. There is value in generating timely evidence on medical treatments using real-world data as a complement to prospective trials.

SUMMARY BOXES

1. *What is already known about this topic?*

- Real-world data is an important source of information about clinical practice, comparative effectiveness and safety, and health outcomes and has the potential to generate timely, pragmatic evidence on medical treatments as a complement to prospective clinical trials.
- Multiple classes of second-line glucose-lowering medications have been approved for the management of type 2 diabetes, with limited evidence of their comparative effectiveness with respect to glycemic control.

2. *What this study adds?*

- We emulated the Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE) randomized clinical trial using data from a U.S. nationwide administrative claims database to identify the strengths and limitations of using real-world data to emulate prospective comparative effectiveness trials, particularly when examining medications in contexts that may not be the standard of care.
- Liraglutide is more effective than glimepiride and sitagliptin at lowering HbA_{1c}, supporting its preferential use when substantial glycemic reduction is needed.
- Advanced causal inference analytic methods applied to observational data can be used to efficiently and effectively emulate clinical trials.

INTRODUCTION

Type 2 diabetes is one of the most common serious chronic health conditions in the U.S. and worldwide, impacting 11.3% (37.3 million) of the U.S. population¹ and 9.3% (463 million) people worldwide.² Moderate glycemic control, defined by achieving glycosylated hemoglobin (HbA_{1c}) between 7% and 8%, improves microvascular and macrovascular outcomes.^{3,4} Current clinical practice guidelines recommend targeting HbA_{1c} under 7% for most non-pregnant adults.⁵ Timely and appropriate treatment intensification is fundamental to maintaining glycemic control⁶ and preventing complications.⁷⁻¹⁰ Metformin is recommended as the first-line glucose-lowering medication due to its efficacy, tolerability, and low cost.¹¹⁻¹⁴ However, type 2 diabetes is a progressive disease and most patients ultimately require intensification of therapy. Recent population-level estimates suggest that nearly one-third of patients with HbA_{1c} $\geq$ 7% are treated with only one glucose-lowering drug¹⁵ and as such would benefit from treatment intensification. Clinical practice guidelines advise that choice of second-line therapy should be informed by clinical and situational considerations specific to each patient, acknowledging the knowledge gaps stemming from the lack of direct comparisons of currently available second-line medications.¹¹⁻¹⁴

The Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE) is a recently completed, but still unpublished, pragmatic, randomized, parallel-arm clinical trial that seeks to address this knowledge gap by comparing four second-line glucose-lowering medications among adults with moderately uncontrolled type 2 diabetes on metformin monotherapy.^{16,17} The drugs represent four medication classes: glimepiride (sulfonylurea), sitagliptin (dipeptidyl-peptidase 4 inhibitor [DPP-4i]), liraglutide (glucagon-like peptide-1 receptor agonist [GLP-1RA]), and insulin glargine (basal analog insulin). GRADE was designed

(2008) and launched (July 2013) prior to U.S. Food and Drug Administration (FDA) approval of sodium-glucose cotransporter-2 inhibitors (SGLT2i) and several cardiovascular outcomes trials that demonstrated reduction in atherosclerotic cardiovascular (ASCVD) and kidney disease outcomes with GLP-1RA use, and in heart failure and kidney disease outcomes with SGLT2i use. This highlights a key limitation of large prospective randomized controlled trials (RCTs): they are time-consuming to conduct, potentially hindering their ability to answer questions in a clinically meaningful time frame. Thus, there is value in efficiently generating timely evidence on medical treatments using observational research methods applied to real-world data as a complement to prospective trials.

Advances in the quantity, quality, and granularity of real-world data, combined with improvements in statistical methods used to account for confounding, treatment allocation bias, and time-related bias, have provided opportunities to use large-scale real-world data to inform our understanding of drug effectiveness and safety. Ideally, studies using real-world data would be conducted prior to the publication of RCT results, thereby minimizing potential biases that could be introduced by trying to replicate known RCT results. As an illustrative test case of the opportunities and limitations of using observational research methods to emulate RCTs, and building on parallel analyses emulating the PRONOUNCE trial,¹⁸ we used claims and laboratory results data from OptumLabs[®] Data Warehouse (OLDW), a de-identified national dataset of privately insured and Medicare Advantage beneficiaries, to emulate the GRADE trial. We used published^{16 19} and publicly available¹⁷ information on GRADE's study design to emulate the methods and anticipated results as closely as possible, with the goal of directly comparing the effectiveness of glimepiride, sitagliptin, liraglutide, and insulin glargine with respect to achieving and maintaining HbA_{1c} below 7.0% among adults with type 2 diabetes and HbA_{1c} 6.8-8.5% on

metformin monotherapy. We also examined the secondary metabolic, microvascular, macrovascular, and safety endpoints planned in GRADE as feasible using the data available within OLDW. This work therefore has two complementary objectives. First – a clinical objective – to examine four second-line glucose-lowering medications with respect to lowering and/or maintaining HbA_{1c} below 7.0%, filling an important clinical knowledge gap with respect to the comparative effectiveness of these commonly used and guideline-recommended drug classes. Second – a methodologic objective – to ascertain whether routinely available claims data can be used to emulate a prospective randomized clinical trial ahead of its publication, filling important methodologic and regulatory policy needs with respect to the use of real-world data to predict clinical trial results.

METHODS

Study design: We retrospectively analyzed medical and pharmacy claims data from OLDW, a nationwide de-identified claims dataset comprised of enrollees in commercial and Medicare Advantage health plans. Private health plan beneficiaries are working-age adults and their dependents. Medicare Advantage plans are Medicare-approved plans offered by private companies to beneficiaries who are eligible for Medicare (e.g., adults 65 years and older, individuals with disability, patients with end-stage kidney disease) as a private alternative to Original Medicare. Just like commercial insurance, Medicare Advantage plans typically bundle medical and pharmacy coverage. OLDW contains longitudinal health information on enrollees in these health plans, representing a diverse mixture of ages, ethnicities, and geographic regions across the U.S.^{20 21} All study data are de-identified consistent with HIPAA expert de-identification determination. The study was exempt from Mayo Clinic Institutional Review

Board review and is reported according to the Reporting of studies Conducted using Observational Routinely-collected Data (RECORD) reporting guideline.²²

Data Sharing Statement: OLDW data are available for research through a virtual data warehouse. The authors are not able to distribute the data. Study protocol, code sets, and statistical analysis plan are available online.²³

Study population: We first assembled a cohort of adults (≥ 18 years) who first started glimepiride, sitagliptin, liraglutide, or insulin glargine between January 25, 2010 (date of liraglutide approval by the FDA; remaining study drugs were approved earlier) and June 30, 2019 (**Figure S1 in the Supplement**). The index date was set to the date of the first claim for the study drug. Patients who started ≥ 2 study drugs on the index date were excluded. Patients were required to be adherent to metformin for ≥ 8 weeks prior to that first fill date. This was established by identifying all metformin fills prior to the index date, establishing continuous treatment episodes based on prescription fill dates and the days' supply for each fill (allowing up to 30-day gap between fills), and requiring that the last metformin treatment episode prior to the index date be at least 8 weeks. To ensure consistent and adequate capture of baseline comorbidities and treatment data, patients were required to have 6 months of continuous enrollment with medical and pharmacy coverage before the index date. Patients with fills for any glucose-lowering medications other than metformin during the baseline period were excluded. Patients with type 1 diabetes, defined using International Classification of Diseases (ICD) codes were excluded. Patients were further required to have valid demographic (age, sex, region) data and HbA_{1c} results both within 3 months prior to the index date ("baseline HbA_{1c}") and during

follow-up. Next, eligibility criteria for the GRADE trial^{16 17 19} were adapted and applied to beneficiaries included in OLDW as detailed in **Table S1**. All relevant diagnosis codes and medications are summarized in **Tables S2-S3**. All eligible patients in OLDW were included in the cohort.

Outcomes: The primary outcome was time to primary metabolic failure of the assigned treatment, calculated as days to $HbA_{1c} \geq 7.0\%$ while treated with the assigned medication, with the period of eligibility starting at month 3 after the index date (analogous to the first quarterly HbA_{1c} assessment in GRADE). Deviating from the GRADE protocol, we did not require a confirmatory HbA_{1c} due to variation in real-world HbA_{1c} testing intervals. To assess for potential bias in outcome ascertainment as the result of different frequencies of HbA_{1c} testing and varying intervals between tests among the treatment groups, we compared these number, frequency, and timing of available HbA_{1c} test results and found no difference between the groups (**Table S4**). Because testing frequency is guided by baseline HbA_{1c} , we also examined intervals between sequential HbA_{1c} tests stratified by baseline HbA_{1c} , and found no differences between the treatment groups (**Table S5**).

Secondary metabolic, cardiovascular, and microvascular outcomes were analyzed as specified in the GRADE statistical analysis plan (SAP)¹⁷ if they were feasible to ascertain using claims data (**Table S6**). Patients were followed until they experienced the outcome of interest, anticipated follow-up duration of the trial (7 years), end of the study period (July 31, 2019), end of insurance coverage, or death. For outcomes observed “while being treated assigned regimen,” we followed patients until they discontinued the assigned medication (defined as not refilling a medication after 30 days of the end of last treatment episode), with the goal of emulating the

definitions of these outcomes in the GRADE trial (i.e., “while treated with originally assigned medications”)¹⁶.

Independent Variables: Patient individual-level age, sex, race/ethnicity, and annual household income were identified from OLDW enrollment files at the time of the index date. Detailed description of the source data for these variables is available in the **Supplemental Methods**. Comorbidities (ascertained from all claims during 6 months preceding the index date) included retinopathy, nephropathy, neuropathy, coronary artery disease, cerebrovascular disease, peripheral vascular disease, heart failure, and prior severe hypoglycemia and hyperglycemia as detailed in **Table S2**. Specialties of treating physicians were categorized as primary care, endocrinology, cardiology, nephrology, other, and unknown. Baseline medications, included as surrogates for complications burden, were identified from fills in the six months preceding the index date (**Table S3**).

Patient and Public Involvement: Patients were not involved in the design, conduct, or dissemination of this study. However, this study was informed by the need to 1) identify preferred glucose-lowering treatment strategies in the absence of direct comparisons across the examined drugs; and 2) examine whether and how data collected in the process of routine patient care can be used to emulate prospective clinical trials. Because this study seeks to inform drug regulatory policy and procedures, investigators from the U.S. FDA contributed to the design of the study and interpretation of study findings; they are included as co-authors on this publication.

Statistical analysis: Inverse probability of treatment weighting (IPTW) were used to balance the differences in baseline characteristics among the treatment groups. Propensity score (PS) were used as probability of treatment; these PS weights were estimated using Generalized Boosted Models (GBM) including the baseline variables presented in **Table 1**. GBM involves an iterative process with multiple regression trees to capture complex and nonlinear relationships between treatment assignments and the pretreatment covariates, resulting in the PS model leading to the best balance among the treatment groups.²⁴ Additional detail on the models is provided in Supp. Methods. Stabilized weights with multiple treatments were calculated by dividing the marginal probability of treatment by the PS of treatment received.²⁵ The distribution of weights is illustrated in **Figure S2**. Standardized mean differences (SMD) were used to assess the balance of covariates after weighting; $SMD \leq 0.1$ was considered a good balance and $SMD \leq 0.2$ was considered acceptable.²⁶ Prior to evaluation of the outcomes, weighted sample sizes and ability to account for baseline confounding were examined to determine the feasibility of including each treatment group.

The cumulative incidence of the primary (time to first $HbA_{1c} \geq 7.0$) and secondary (time to first $HbA_{1c} > 7.5\%$) metabolic failures within each treatment arm was estimated with the IPTW Kaplan-Meier method. PS weighted Cox proportional hazards (PH) regression models adjusted by baseline HbA_{1c} values were used to compare the outcomes between treatment groups. The at-risk time for the PH model was set as 3 months after the index date as the primary outcome can be only observed starting at 3rd month. Results were presented as median times to metabolic failure and expected proportions of patients experiencing metabolic failure at 1 and 2 years. All pairwise comparisons between the treatment groups were estimated and the Holm method was applied to adjust the p-values for multiple testing. The proportional hazards assumption was

tested using Schoenfeld residuals. Similar analyses were performed for other time-to-event outcomes. The at-risk start time for modeling secondary metabolic, cardiovascular, and microvascular disease outcomes was set at the study index date. Repeated measures HbA_{1c} trends by treatment group were estimated by using the IPTW mean HbA_{1c} results by treatment group in 3-month time intervals. The follow-up time by treatment arm was estimated using the same propensity score weights as the primary analysis and the IPTW Kaplan-Meier method for the censoring distribution.²⁷

All primary analyses were conducted using the “per-protocol” censoring approach for the primary outcome and for the secondary outcomes of secondary metabolic failure and insulin initiation, censoring at the time of treatment drug discontinuation, disenrollment from the health plan, end of study period, or death, whichever came first (**Figure S3**). Time “on treatment” for each drug was determined by calculating continuous coverage episodes based on available fills, same as for baseline metformin therapy. Remaining secondary outcomes were analyzed using the “intention-to-treat” censoring approach, censoring the patient at the time of health plan disenrollment, end of study, or death, which ever came first. $P < 0.05$ was considered statistically significant for all 2-sided tests. All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC) and R version 4.0.2.(R Foundation).

Subgroup analyses: A priori-defined subgroup analyses were performed as a function of baseline HbA_{1c} (<7.0% vs. ≥7.0%), age group (<65 years, ≥65 years), sex (male vs. female) and race/ethnicity (White, Black, Hispanic, Asian).

Sensitivity analyses: First, to examine the comparative effectiveness of study medications while treated only with them and not with any other medication, accounting for real-world treatment practices, we repeated all analyses using the “as-treated” censoring approach, censoring at the time a new medication class was added, the assigned medication was discontinued, health plan disenrollment, end of study, or death, which ever came first (**Figure S3**). Second, we assessed residual confounding by testing a falsification end point that is unlikely to be associated with the studied medications: diagnosis of pneumonia (**Table S2**) during the follow-up period.

RESULTS

Study Population

We identified 18,365 adults with type 2 diabetes who started glimepiride, 12,818 who started sitagliptin, 5,021 who started liraglutide, and 5,659 who started insulin glargine and had the required baseline enrollment and available HbA_{1c} results (**Figure S1**). GRADE eligibility criteria were met by 19.7% of these patients, ranging from 4.4% of glargine-treated patients to 23.5% of glimepiride-treated patients. The most prevalent reasons for ineligibility (**Table S7**) were HbA_{1c} outside the prespecified range (ranging from 51.7% of sitagliptin-treated patients to 81.1% of glargine-treated patients) and not being treated with metformin monotherapy at the time of study drug initiation (ranging from 43.5% of glimepiride-treated patients to 68.5% of glargine-treated patients). The final cohort was comprised of 4318 patients in the glimepiride arm, 2993 patients in the sitagliptin arm, 690 patients in the liraglutide arm, and 251 patients in the glargine arm (see **Table S8** for all included drugs).

Baseline patient characteristics prior to weighting are shown in **Table S9**. There were significant differences (largest SMD >0.2) in patient age, race/ethnicity, annual household

income, and prescribing physician specialty across the four treatment groups. Patients in the liraglutide arm were more likely to be younger, White, higher income, and treated by an endocrinologist than patients in the other treatment arms. Patients in the glargine arm were most likely to be low income and had the highest prevalence of all examined comorbidities.

The glargine arm was excluded from all analyses because of small sample size (N=251, weighted N=179) and inability to achieve good confounder control after weighting. After the glargine group was dropped, the PS model was estimated on the remaining 3 groups. After weighting, mean patient ages were 62.0 (SD, 11.1) years in the glimepiride arm, 62.0 (SD, 11.0) in the sitagliptin arm, and 60.5 (SD, 10.4) in the liraglutide arm (**Table 1**). Women comprised 48.2%, 49.1%, and 50.5% of the treatment arms, respectively. White patients comprised 64.7%, 64.2%, and 65.8% of the treatment arms, respectively. Mean baseline HbA_{1c} levels were 7.63% (SD, 0.48), 7.61% (SD, 0.47), and 7.60% (SD, 0.48), respectively. Pairwise SMDs for all baseline covariates are presented in **Table S10**; all SMDs were <0.2.

Primary Metabolic Failure (HbA_{1c} ≥7.0%)

Median follow-up until per-protocol censoring was 238 (95% CI, 226-255) days in the glimepiride arm, 124 (95% CI, 100-150) days in the liraglutide arm, and 186 (95% CI, 179-201) days in the sitagliptin arm (**Figure S4**). Mean HbA_{1c} decreased most in the liraglutide arm and least in the sitagliptin arm, with differences most pronounced between months 3 and 6 of treatment (**Figure 1A**). The median times to primary metabolic failure were 442 (95% CI, 394-480) days in the glimepiride arm, 764 (95% CI, 741-NA) days in the liraglutide arm, and 427 (95% CI, 380-483) days in the sitagliptin arm (**Figure 1B**). Liraglutide was associated with lower risk of primary metabolic failure compared to glimepiride (HR 0.57; 95% CI, 0.43-0.75)

and sitagliptin (HR 0.55; 95% CI, 0.41-0.73); **Table 2**. No significant difference was observed between sitagliptin and glimepiride (HR 1.03; 95% CI, 0.94-1.13). By 1-year, the estimated rates of primary metabolic failure were 0.28 (95% CI, 0.19-0.36) in the liraglutide arm, 0.44 (95% CI, 0.42-0.46) in the glimepiride arm, and 0.46 (95% CI, 0.43-0.48) in the sitagliptin arm (**Table 3**). The difference in event rates persisted at 2 years.

Secondary Metabolic Failure ($HbA_{1c} >7.5\%$)

Time to secondary metabolic failure was longest in the liraglutide arm (**Figure S5**). Liraglutide was associated with lower risk of secondary metabolic failure compared to glimepiride (HR 0.61; 95% CI [0.43, 0.87]) and sitagliptin (HR 0.59; 95% CI [0.41, 0.85]); (**Table 2**). By 1-year, the estimated rates of secondary metabolic failure were 0.11 (95% CI, 0.06-0.17) in the liraglutide arm, 0.20 (95% CI, 0.19-0.22) in the glimepiride arm, and 0.22 (95% CI, 0.19-0.24) in the sitagliptin arm (**Table 3**). The difference in event rates persisted at 2 years.

Other Secondary Outcomes

Insulin was started by 84 (2.0%) patients in the glimepiride arm, 11 (1.9%) patients in the liraglutide arm, and 50 (1.8%) patients in the sitagliptin arm, with no significant difference among the three groups (**Table 2**). Overall, 37 patients experienced emergency department visits or hospitalizations for hypoglycemia during the study period, including <11 in the liraglutide and sitagliptin arms, precluding statistical analyses.

Heart failure, ESKD, pancreatitis, pancreatic/thyroid cancer, and all-cause mortality could not be analyzed due to <11 events in all treatment groups (event rates are presented in **Table S11**). There were no statistically significant differences between groups for MACE,

retinopathy, neuropathy, other cardiovascular events, cancer, and all-cause hospitalizations (Table S12).

Subgroup Analyses

Liraglutide was associated with lower risk of primary metabolic failure compared with glimepiride (HR 0.59; 95% CI, 0.44-0.78) and sitagliptin (HR 0.58; 95% CI, 0.43-0.79) among patients with baseline HbA_{1c} ≥7.0%. There were no significant differences among the treatment groups in patients with baseline HbA_{1c} <7.0% (Table S13). Liraglutide was associated with lower risk of primary metabolic failure compared with glimepiride (HR 0.54; 95% CI, 0.42-0.71) and sitagliptin (HR 0.58; 95% CI, 0.44-0.77) among patients <65 years old. There were no significant differences among groups in patients ≥65 years of age. Liraglutide was also associated with lower risks of primary metabolic failure than glimepiride and sitagliptin in women, but not in men, and in White and Hispanic patients, but not in Black or Asian patients. Findings were similar for secondary metabolic failure (Table S14).

Sensitivity Analyses

Another glucose-lowering medication was added prior to discontinuation of the assigned treatment in 434 (10%) patients in the glimepiride arm, 268 (39%) in the liraglutide arm, and 429 (14%) in the sitagliptin arm. Sensitivity analyses using the as-treated censor approach were consistent with the primary analyses (Figure S6, Table S15). There were no significant differences among the treatment groups in the pneumonia falsification endpoint (Table S16).

DISCUSSION

Our emulation of the GRADE trial using real-world data from an administrative claims database revealed that liraglutide was significantly more effective at maintaining glycemic control, defined by time to $\text{HbA}_{1c} \geq 7.0\%$ (primary metabolic failure) and $\text{HbA}_{1c} > 7.5\%$ (secondary metabolic failure) than either glimepiride or sitagliptin. These differences are clinically meaningful, with over 40% more patients in control of their HbA_{1c} when treated with liraglutide than when treated with glimepiride or sitagliptin. We were unable to include insulin glargine in the comparisons because of the small number of glargine-treated patients meeting GRADE eligibility criteria. This was not surprising as treatment with basal insulin in the clinical context examined by GRADE is outside the standard of care and mainstream practice. Additionally, the analytic framework implemented in this work demonstrates that real-world data may be an important complement to prospective trials, allowing for efficient and timely examination of pressing clinical questions and inquiries of comparative effectiveness and safety.

Our efforts to emulate all specifications of the GRADE trial were hindered by the fact that study conditions are not adequately represented in real-world practice as they are not supported by clinical practice guidelines. While all four study drugs were frequently used by the OLDW population, 80% of adults starting these medications had to be excluded because they did not meet the prespecified GRADE eligibility criteria. Nevertheless, this proportion of included patients is still higher than the 9.1% generalizability estimated by the GRADE study team when compared to the overall U.S. population with diabetes.¹⁹ The majority of patients (58.6% overall) were excluded due to not meeting baseline HbA_{1c} level requirements, including 81.1% of glargine-initiators, 71.8% of liraglutide-initiators, 52.8% of glimepiride initiators, and 51.7% of sitagliptin-initiators. According to current guidelines, the target HbA_{1c} for most non-pregnant adults is 7.0%, such that treatment intensification would not be warranted for some participants.

Initiation of insulin, in particular, is advised when HbA_{1c} is above 9-10%,^{14 28} so starting glargine as a second-line drug at HbA_{1c} levels below 8.5% would not be consistent with the standard of care^{14 28} or contemporary practice.²⁹⁻³¹ The fact that most patients treated with the examined medications in clinical practice are not represented in the study population raises concerns about the utility and generalizability of GRADE study findings and its impact on diabetes management, underscoring the important complementary insights that can be gleaned from analyses of real-world data (which can be designed to use more pragmatic and generalizable eligibility criteria) as adjuncts to RCTs.

We met our objective to conduct all analyses prior to publication of GRADE trial findings, and it will be important to ultimately compare our findings to those in GRADE. The greater effectiveness of liraglutide compared to both glimepiride and sitagliptin is consistent with prior studies.^{30 32-34} Additionally, subgroup analyses demonstrating greater effectiveness of liraglutide among patients with elevated baseline HbA_{1c} and in younger patients, generated important hypotheses regarding the optimal use of liraglutide (and potentially other GLP-1RA) in clinical practice to be explored in future research. When the GRADE trial was conceived, drugs' ability to lower HbA_{1c} was at the forefront of clinical decision making when choosing glucose-lowering therapy. Similarly, the SGLT2i class of glucose-lowering medications had not yet been incorporated into practice and therefore was excluded as a comparator therapy when the GRADE trial was conceived and designed. In addition, contemporary clinical practice guidelines increasingly focus on the impacts of glucose-lowering therapies on hard outcomes that are important to patients such macrovascular and microvascular complications and death, beyond HbA_{1c}.³⁵ Indeed, most recent clinical practice guidelines recommend consideration of GLP-1RA and SGLT2i even as first-line therapies and independent of the HbA_{1c} level among patients at

high risk for ASCVD, kidney disease, and heart failure.¹⁴ For these outcomes, there is robust evidence favoring liraglutide (of the drug classes examined) in patients at high risk for ASCVD,^{36 37} further underscoring its advantage. It will be important, in future research, to compare the effectiveness of glycemic control achieved by GLP-1RA with that of SGLT2i, as SGLT2i are similarly recommended for patients at high risk for cardiovascular disease, kidney disease, and heart failure.¹⁴

Our study is strengthened by application of advanced analytic methods that account for measured differences between treatment arms that otherwise confound analyses and preclude causal inference. The GBM-based models for the PS are more flexible and less sensitive to model misspecification compared to logistic regression. The large and diverse patient population within OLDW made emulation efforts uniquely possible despite the narrow eligibility criteria specified by GRADE.

Limitations

Despite rigorous causal inference analytic methods, observational studies are inevitably subject to residual confounding. For the metabolic endpoints, there was evidence of non-proportional hazards, which makes the single summary hazard ratio calculated from the Cox proportional hazards an imperfect estimate for the time-varying risk. However, with the goal of emulating the GRADE trial, where the statistical analysis plan was to estimate single summary hazard ratios, we report the same estimate in the emulation. Nevertheless, we were unable to operationalize every component of GRADE's eligibility criteria and end points. For example, we did not require confirmatory HbA_{1c} results to meet the metabolic endpoints and were not able to maintain the same standard timeframe for HbA_{1c} ascertainment as specified in GRADE.

Additionally, while the GRADE trial analyses were conducted using the ITT principle, we a priori chose to use per-protocol analysis for the metabolic endpoints because in the absence of randomization, reasons for changing a treatment typically depend on post-initiation factors that could confound the association between the treatment group and the outcome. While advanced statistical methods can account for post-baseline differences between groups in key characteristics, these methods require accurate estimation of the reasons to stop or change treatment, and such estimation is not feasible in this setting using claims data. Duration of follow-up was also different among the treatment arms, which is unavoidable when studying real-world practice patterns. In particular, a higher proportion of patients initiating liraglutide filled only one cycle of treatment before either switching to a different treatment or not refilling their prescription, potentially due to poor tolerability, subcutaneous administration, or high cost.

Not all patients with claims data in OLDW have available laboratory data, as results are available for a subset of patients based on data sharing agreements between OptumLabs and commercial laboratories. However, laboratory results availability is independent of treatment regimen, and we do not expect it to bias our analyses. The schedule of HbA_{1c} testing in real-world practice is contingent on the patient's current HbA_{1c} level, the clinician's anticipation of changing HbA_{1c} levels, and the patient's ability to access care. This may have confounded study results by delaying the time to HbA_{1c} reassessment and reaching the study endpoint in patients with low baseline HbA_{1c} or with barriers to care. Our evaluation could not account for inclusion and exclusion criteria that could not be operationalized using claims data, including medications obtained without insurance coverage (e.g., obtained through a low-cost generic program,³⁸ patient assistance program, or a sample), comorbidities that were not coded and billed in a clinical encounter, and family history information. However, prior studies found the likely

number of glucose-lowering medications missing from claims to be low.³⁹ Finally, the study cohort was comprised of patients with private and Medicare Advantage health plans, such that results may not fully generalize to patients with public health plans or those without insurance coverage.

CONCLUSIONS

Better understanding of the comparative effectiveness and safety of second-line glucose-lowering medications is urgently needed to inform shared decision-making in diabetes. Analytic methods such as those implemented in this study, and in the parallel emulation of PRONOUNCE,¹⁸ can be leveraged for more timely evaluations of drug effectiveness and safety as long as the treatments being considered are already being used in clinical practice. Indeed, work is currently under way to examine the comparative effectiveness of sulfonylurea, GLP-1RA, DPP-4i, and SGLT2i drugs with respect to ASCVD and other hard outcomes among patients at moderate risk for ASCVD using observational data from real-world practice.⁴⁰ Ultimately, the population included in this study and our findings should be compared to those of the GRADE trial, once published in peer-reviewed literature, to assess the fidelity and generalizability of results and to deepen our understanding of the use of real-world data to emulate clinical trials.

ACKNOWLEDGEMENTS

Funding: This publication is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance (Center of Excellence in Regulatory Science and Innovation grant to Yale University and Mayo Clinic

U01FD005938) totaling \$250,000 with 100 percent funded by FDA/HHS. Dr. McCoy is also supported by the National Institute of Health (NIH) National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant number K23DK114497.

Role of the Sponsor: The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, NIH, or the U.S. Government.

Conflicts of Interest: In the last 36 months, Dr. McCoy also received support from NIDDK, PCORI, and AARP[®]. She has also served as a consultant to Emmi[®] on the development of patient education materials related to prediabetes and diabetes. Dr. Dhruva reports receiving funding support from the National Heart, Lung, and Blood Institute (NHLBI) of the NIH, National Evaluation System for health Technology Coordinating Center (NESTcc), the Greenwall Foundation, and Arnold Ventures. Dr. Wallach has received research support from the National Institute on Alcohol Abuse and Alcoholism of the NIH, and the Collaboration for Research Integrity and Transparency (CRIT) at Yale University from the Laura and John Arnold Foundation. Dr Herrin works under contract to the Centers for Medicare and Medicaid Services on the development and evaluation of measures of provider quality. In the past 36 months, Dr. Shah has received support through Mayo Clinic from the Food and Drug Administration to establish the Yale-Mayo Clinic CERSI program (U01FD005938); the Centers of Medicare and Medicaid innovation under the Transforming Clinical practice Initiative; the Agency for Healthcare Research and Quality (R01HS025164, R01HS025402, R03HS025517, and K12HS026379); the National Heart, Lung and Blood Institute of the US National Institutes of Health (R56HL130496, R01HL131535, and R01HL151662); the National Science Foundation; and the Patient Centered Outcomes Research Institute to Develop a Clinical Data Research Network (LHSNet). In the past 36 months, Dr. Ross received research support through Yale

University from the Laura and John Arnold Foundation for the Collaboration for Research Integrity and Transparency (CRIT) at Yale, from Medtronic, Inc. and the Food and Drug Administration (FDA) to develop methods for postmarket surveillance of medical devices (U01FD004585) and from the Centers of Medicare and Medicaid Services (CMS) to develop and maintain performance measures that are used for public reporting (HHSM-500-2013-13018I); Dr. Ross currently receives research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing, from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology (NEST), from the Food and Drug Administration for the Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938); from the Agency for Healthcare Research and Quality (R01HS022882), from the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) (R01HS025164, R01HL144644), and from the Laura and John Arnold Foundation to establish the Good Pharma Scorecard at Bioethics International.

Data Sharing: This study was conducted using de-identified claims data from OptumLabs Data Warehouse. Raw data are not publicly available. The study protocol is available online.²³

REFERENCES

1. CDC. National Diabetes Statistics Report Atlanta, Georgia, U.S.A.: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2021 [updated December 29, 2021; cited 2022 May 17]. Available from: <https://www.cdc.gov/diabetes/data/statistics-report/> accessed May 17 2022.
2. Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9(th) edition. *Diabetes Res Clin Pract* 2019;157:107843. doi: 10.1016/j.diabres.2019.107843 [published Online First: 2019/09/14]
3. Hemmingsen B, Lund SS, Gluud C, et al. Intensive glycaemic control for patients with type 2 diabetes: systematic review with meta-analysis and trial sequential analysis of randomised clinical trials. *BMJ* 2011;343:d6898. doi: 10.1136/bmj.d6898
4. Rodriguez-Gutierrez R, Gonzalez-Gonzalez JG, Zuniga-Hernandez JA, et al. Benefits and harms of intensive glycemic control in patients with type 2 diabetes. *BMJ* 2019;367:l5887. doi: 10.1136/bmj.l5887 [published Online First: 2019/11/07]

5. American Diabetes Association Standards of Medical Care in Diabetes—2020. Section 6. Glycemic Targets. *Diabetes Care* 2020;43(Supplement 1):S66-S76. doi: 10.2337/dc20-S006
6. Schmittiel JA, Uratsu CS, Karter AJ, et al. Why don't diabetes patients achieve recommended risk factor targets? Poor adherence versus lack of treatment intensification. *J Gen Intern Med* 2008;23(5):588-94. doi: 10.1007/s11606-008-0554-8 [published Online First: 2008/03/05]
7. Shichiri M, Kishikawa H, Ohkubo Y, et al. Long-term results of the Kumamoto Study on optimal diabetes control in type 2 diabetic patients. *Diabetes Care* 2000;23 Suppl 2:B21-9. [published Online First: 2000/06/22]
8. Turner RC, Cull CA, Frighi V, et al. Glycemic control with diet, sulfonylurea, metformin, or insulin in patients with type 2 diabetes mellitus: Progressive requirement for multiple therapies (ukpds 49). *JAMA* 1999;281(21):2005-12. doi: 10.1001/jama.281.21.2005
9. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* 2000;321(7258):405-12. [published Online First: 2000/08/11]
10. Holman RR, Paul SK, Bethel MA, et al. 10-year follow-up of intensive glucose control in type 2 diabetes. *New England Journal of Medicine* 2008;359(15):1577-89. doi: 10.1056/NEJMoa0806470 [published Online First: 2008/09/12]
11. NICE. National Institute for Health and Care Excellence Pathways: Managing Blood Glucose in Adults with Type 2 Diabetes: National Institute for Health and Care Excellence; 2019 [updated March 26, 2019; cited 2019 April 23]. Available from: <https://pathways.nice.org.uk/pathways/type-2-diabetes-in-adults> accessed April 23 2019.
12. Garber AJ, Abrahamson MJ, Barzilay JI, et al. Consensus Statement By The American Association Of Clinical Endocrinologists And American College Of Endocrinology On The Comprehensive Type 2 Diabetes Management Algorithm – 2019 Executive Summary. *Endocrine Practice* 2019;25(1):69-100. doi: 10.4158/cs-2018-0535
13. Conlin PR, Colburn J, Aron D, et al. Synopsis of the 2017 U.S. Department of Veterans Affairs/U.S. Department of Defense Clinical Practice Guideline: Management of type 2 diabetes mellitus. *Annals of Internal Medicine* 2017;167(9):655-63. doi: 10.7326/M17-1362
14. American Diabetes Association Standards of Medical Care in Diabetes—2022. Section 9. Pharmacologic Approaches to Glycemic Treatment. *Diabetes Care* 2021;45(Supplement_1):S125-S43. doi: 10.2337/dc22-S009
15. Fang M, Wang D, Coresh J, et al. Trends in Diabetes Treatment and Control in U.S. Adults, 1999–2018. *New England Journal of Medicine* 2021;384(23):2219-28. doi: 10.1056/NEJMsa2032271
16. Nathan DM, Buse JB, Kahn SE, et al. Rationale and design of the glycemia reduction approaches in diabetes: a comparative effectiveness study (GRADE). *Diabetes Care* 2013;36(8):2254-61. doi: 10.2337/dc13-0356 [published Online First: 2013/05/22]
17. ClinicalTrials.gov. A Comparative Effectiveness Study of Major Glycemia-lowering Medications for Treatment of Type 2 Diabetes (GRADE). ClinicalTrials.gov Identifier: NCT01794143; U.S. National Library of Medicine. National Institutes of Health. Department of Health and Human Services; 2021 [updated June 30, 2021; cited 2021 September 7]. Available from: <https://clinicaltrials.gov/ct2/show/NCT01794143> accessed September 7 2021.
18. Wallach JD, Deng Y, McCoy RG, et al. Real-world Cardiovascular Outcomes Associated With Degarelix vs Leuprolide for Prostate Cancer Treatment. *JAMA Netw Open* 2021;4(10):e2130587. doi: 10.1001/jamanetworkopen.2021.30587 [published Online First: 2021/10/23]
19. Wexler DJ, Krause-Steinrauf H, Crandall JP, et al. Baseline Characteristics of Randomized Participants in the Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE). *Diabetes Care* 2019;42(11):2098-107. doi: 10.2337/dc19-0901 [published Online First: 2019/08/09]
20. Wallace PJ, Shah ND, Dennen T, et al. Optum Labs: Building A Novel Node In The Learning Health Care System. *Health Aff (Millwood)* 2014;33(7):1187-94. doi: 10.1377/hlthaff.2014.0038

21. OptumLabs. OptumLabs and OptumLabs Data Warehouse (OLDW) Descriptions and Citation. Cambridge, MA: n.p., May 2019. PDF. Reproduced with permission from OptumLabs., 2019.
22. Benchimol EI, Smeeth L, Guttman A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12(10):e1001885. doi: 10.1371/journal.pmed.1001885 [published Online First: 2015/10/07]
23. Yale University-Mayo Clinic CERSI 2021 [cited 2021 September 29]. Available from: https://medicine.yale.edu/core/current_projects/cersi/research/ accessed September 29 2021.
24. McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* 2013;32(19):3388-414. doi: <https://doi.org/10.1002/sim.5753>
25. Xu S, Ross C, Raebel MA, et al. Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals. *Value in Health* 2010;13(2):273-77. doi: <https://doi.org/10.1111/j.1524-4733.2009.00671.x>
26. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 2009;28(25):3083-107. doi: <https://doi.org/10.1002/sim.3697>
27. Korn EL. Censoring distributions as a measure of follow-up in survival analysis. *Statistics in Medicine* 1986;5(3):255-60. doi: <https://doi.org/10.1002/sim.4780050306>
28. American Diabetes Association Standards of Medical Care in Diabetes—2022. Section 6. Glycemic Targets. *Diabetes Care* 2021;45(Supplement_1):S83-S96. doi: 10.2337/dc22-S006
29. Liss DT, Kang RH, Lancki N, et al. Costs for commercially insured adults prescribed second-line diabetes medications. *Am J Manag Care* 2021;27(3):e72-e79. doi: 10.37765/ajmc.2021.88601 [published Online First: 2021/03/16]
30. Khunti K, Charbonnel B, Cooper A, et al. Associations between second-line glucose-lowering combination therapies with metformin and HbA1c, body weight, quality of life, hypoglycaemic events and glucose-lowering treatment intensification: The DISCOVER study. *Diabetes Obes Metab* 2021 doi: 10.1111/dom.14400 [published Online First: 2021/04/15]
31. Nicolucci A, Charbonnel B, Gomes MB, et al. Treatment patterns and associated factors in 14 668 people with type 2 diabetes initiating a second-line therapy: Results from the global DISCOVER study programme. *Diabetes Obes Metab* 2019;21(11):2474-85. doi: 10.1111/dom.13830 [published Online First: 2019/07/13]
32. Tzapas A, Avgerinos I, Karagiannis T, et al. Comparative Effectiveness of Glucose-Lowering Drugs for Type 2 Diabetes: A Systematic Review and Network Meta-analysis. *Ann Intern Med* 2020;173(4):278-86. doi: 10.7326/m20-0864 [published Online First: 2020/07/01]
33. Tran S, Retnakaran R, Zinman B, et al. Efficacy of glucagon-like peptide-1 receptor agonists compared to dipeptidyl peptidase-4 inhibitors for the management of type 2 diabetes: A meta-analysis of randomized clinical trials. *Diabetes Obes Metab* 2018;20 Suppl 1:68-76. doi: 10.1111/dom.13137 [published Online First: 2018/01/25]
34. Palmer SC, Mavridis D, Nicolucci A, et al. Comparison of Clinical Outcomes and Adverse Events Associated With Glucose-Lowering Drugs in Patients With Type 2 Diabetes: A Meta-analysis. *Jama* 2016;316(3):313-24. doi: 10.1001/jama.2016.9400 [published Online First: 2016/07/21]
35. Rodriguez-Gutierrez R, McCoy RG. Measuring What Matters in Diabetes. *JAMA* 2019 doi: 10.1001/jama.2019.4310 [published Online First: 2019/04/16]
36. Mann JFE, Orsted DD, Brown-Frandsen K, et al. Liraglutide and Renal Outcomes in Type 2 Diabetes. *N Engl J Med* 2017;377(9):839-48. doi: 10.1056/NEJMoa1616011 [published Online First: 2017/08/31]
37. Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. *N Engl J Med* 2016;375(4):311-22. doi: 10.1056/NEJMoa1603827 [published Online First: 2016/06/14]

38. Pauly NJ, Brown JD. Prevalence of Low-Cost Generic Program Use in a Nationally Representative Cohort of Privately Insured Adults. *J Manag Care Spec Pharm* 2015;21(12):1162-70. doi: 10.18553/jmcp.2015.21.12.1162 [published Online First: 2015/12/19]
39. Rowan CG, Flory J, Gerhard T, et al. Agreement and validity of electronic health record prescribing data relative to pharmacy claims data: A validation study from a US electronic health record database. *Pharmacoepidemiol Drug Saf* 2017;26(8):963-72. doi: 10.1002/pds.4234 [published Online First: 2017/06/14]
40. PCORI. Observational Analyses of Second-Line Pharmacological Agents in Type 2 Diabetes Washington, DC: Patient-Centered Outcomes Research Institute; 2021 [cited 2021 September 7]. Available from: <https://www.pcori.org/funding-opportunities/announcement/observational-analyses-second-line-pharmacological-agents-type-2-diabetes-cycle-2-2020> accessed September 7 2021.

Table 1. Baseline characteristics in weighted cohort. Data are presented as count (percentage), unless otherwise specified. ARB, angiotensin II receptor blockers; ACE, angiotensin-converting enzyme; HbA_{1c}, hemoglobin A_{1c}; SMD, standardized mean difference.

	Glimepiride	Liraglutide	Sitagliptin	Largest SMD
Weighted N	4168	572	2800	
Age, years, mean (SD)	62.0 (11.1)	60.5 (10.4)	61.98 (11.0)	0.14
Sex				0.05
Female	2009 (48.2%)	289 (50.5%)	1374 (49.1%)	
Male	2159 (51.8%)	283 (49.5%)	1427 (50.9%)	
Race/ethnicity				0.09
White	2695 (64.7%)	376 (65.8%)	1798 (64.2%)	
Black	536 (12.9%)	79 (13.7%)	355 (12.7%)	
Hispanic	513 (12.3%)	72 (12.6%)	353 (12.6%)	
Asian	266 (6.4%)	28 (4.9%)	185 (6.6%)	
Unknown	159 (3.8%)	17 (3.0%)	108 (3.9%)	
Annual Household Income				0.10
<\$40,000	989 (23.7%)	123 (21.6%)	647 (23.1%)	
\$40,000 - \$74,999	1145 (27.5%)	175 (30.7%)	766 (27.3%)	
\$75,000 – \$124,999	1153 (27.7%)	164 (28.6%)	772 (27.6%)	
\$125,000 – \$199,999	485 (11.6%)	63 (11.0%)	336 (12.0%)	
≥\$200,000	185 (4.4%)	26 (4.5%)	139 (4.9%)	
Unknown/missing	211 (5.1%)	21 (3.7%)	142 (5.1%)	
Baseline HbA_{1c} mean (SD)	7.6 (0.5)	7.6 (0.5)	7.6 (0.5)	0.06
Baseline HbA_{1c} categories				0.09
6.8-6.9%	336 (8.1%)	59 (10.3%)	233 (8.3%)	
7.0-7.9%	2630 (63.1%)	364 (63.7%)	1786 (63.8%)	
8.0-8.5%	1202 (28.8%)	149 (26.0%)	782 (27.9%)	
Baseline creatinine, mg/dL, mean (SD)*	0.9 (0.2)	0.9 (0.2)	0.9 (0.2)	0.02
Baseline comorbidities				
Nephropathy	363 (8.7%)	55 (9.5%)	229 (8.2%)	0.05
Retinopathy	196 (4.7%)	31 (5.4%)	135 (4.8%)	0.03
Neuropathy	505 (12.1%)	75 (13.2%)	326 (11.7%)	0.05
Hyperglycemia	2 (0.0%)	0	0	0.03
Hypoglycemia	2 (0.0%)	0	0	0.03
Coronary artery disease	355 (8.5%)	53 (9.2%)	224 (8.0%)	0.04
Chronic kidney disease	121 (2.9%)	18 (3.1%)	82 (2.9%)	0.01
Cerebrovascular disease	117 (2.8%)	16 (2.8%)	80 (2.8%)	0.00
Peripheral vascular disease	190 (4.6%)	28 (4.9%)	130 (4.6%)	0.02
Baseline medications				

	Glimepiride	Liraglutide	Sitagliptin	Largest SMD
ACE inhibitor	1819 (43.6%)	250 (43.7%)	1206 (43.1%)	0.01
ARB	1110 (26.6%)	157 (27.4%)	755 (27.0%)	0.02
ACE inhibitor or ARB	2850 (68.4%)	400 (69.9%)	1906 (68.0%)	0.04
Direct oral anticoagulant	57 (1.4%)	9 (1.5%)	40 (1.4%)	0.01
Statin	2853 (68.5%)	374 (65.4%)	1937 (69.2%)	0.08
Non-statin lipid lowering medications	549 (13.2%)	85 (14.9%)	382 (13.6%)	0.05
Warfarin	83 (2.0%)	12 (2.0%)	48 (1.7%)	0.02
Peripheral neuropathy medications	410 (9.8%)	74 (12.9%)	281 (10.0%)	0.09
Specialty of treating physicians				0.10
Primary care	3202 (76.8%)	424 (74.1%)	2145 (76.6%)	
Endocrinology	174 (4.2%)	27 (4.7%)	120 (4.3%)	
Cardiology	28 (0.7%)	2 (0.4%)	19 (0.7%)	
Nephrology	6 (0.1%)	0	5 (0.2%)	
Other	274 (6.6%)	45 (7.9%)	179 (6.4%)	
Unknown	483 (11.6%)	74 (12.9%)	333 (11.9%)	
Year of cohort entry				0.13
2010	237 (5.7%)	29 (5.0%)	167 (6.0%)	
2011	241 (5.8%)	25 (4.4%)	168 (6.0%)	
2012	297 (7.1%)	41 (7.2%)	204 (7.3%)	
2013	388 (9.3%)	65 (11.3%)	268 (9.6%)	
2014	464 (11.1%)	65 (11.4%)	318 (11.4%)	
2015	449 (10.8%)	58 (10.1%)	293 (10.4%)	
2016	575 (13.8%)	75 (13.0%)	380 (13.6%)	
2017	761 (18.3%)	97 (16.9%)	497 (17.8%)	
2018	632 (15.2%)	91 (15.9%)	422 (15.1%)	
2019	126 (3.0%)	27 (4.7%)	84 (3.0%)	

*Baseline creatinine was not included in the propensity score model as there were missing values.

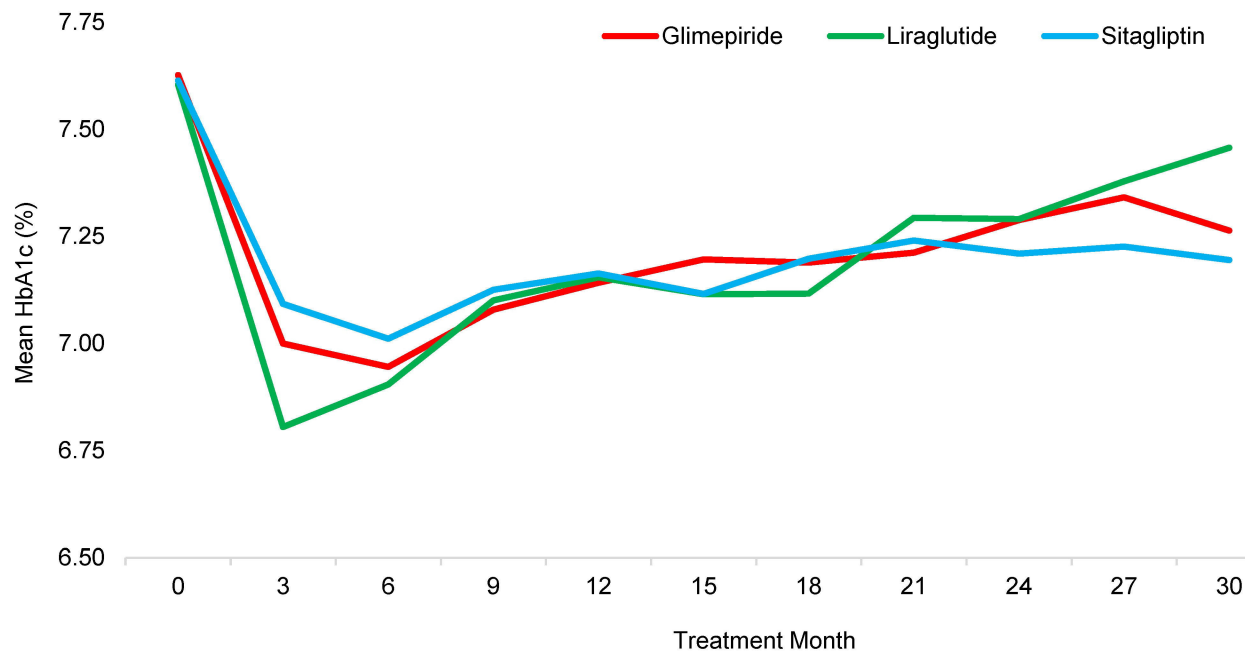
Table 2. Hazard ratios for primary and secondary metabolic outcomes. *Holm-adjusted p-values.

	Hazard Ratio (95% CI)	P-value *
Primary metabolic failure (HbA_{1c} ≥7%)		
Liraglutide vs Glimepiride	0.57 (0.43, 0.75)	<0.001
Sitagliptin vs Glimepiride	1.03 (0.94, 1.13)	0.48
Liraglutide vs Sitagliptin	0.55 (0.41, 0.73)	<0.001
Secondary metabolic failure (HbA_{1c} >7.5%)		
Liraglutide vs Glimepiride	0.61 (0.43, 0.87)	0.01
Sitagliptin vs Glimepiride	1.04 (0.91, 1.18)	0.60
Liraglutide vs Sitagliptin	0.59 (0.41, 0.85)	0.01
Initiating insulin		
Liraglutide vs Glimepiride	1.49 (0.76, 2.93)	0.75
Sitagliptin vs Glimepiride	1.10 (0.77, 1.58)	0.81
Liraglutide vs Sitagliptin	1.35 (0.67, 2.72)	0.81

Table 3. Rates of primary and secondary metabolic failure by treatment arm.

	1-year Event Rate (95% CI)	2-year Event Rate (95% CI)
Primary metabolic failure ($HbA_{1c} \geq 7\%$)		
Glimepiride	0.44 (0.42, 0.46)	0.65 (0.63, 0.68)
Sitagliptin	0.28 (0.19, 0.36)	0.40 (0.29, 0.49)
Liraglutide	0.46 (0.43, 0.48)	0.62 (0.58, 0.65)
Secondary metabolic failure ($HbA_{1c} > 7.5\%$)		
Glimepiride	0.20 (0.19, 0.22)	0.37 (0.34, 0.39)
Sitagliptin	0.11 (0.06, 0.17)	0.17 (0.10, 0.25)
Liraglutide	0.22 (0.19, 0.24)	0.37 (0.33, 0.41)

Figure 1. Glycemic control. (A) Mean HbA_{1c} levels over time. Results are based on observed on-treatment trajectories, with no imputation of missing HbA_{1c} levels. (B) Cumulative risks of primary metabolic failure in propensity score weighted patients.

A**Number of patients at risk**

Treatment month:	0	3	6	9	12	15	18	21	24	27	30
Glimepiride	4340	1454	2449	1967	1823	1593	1389	1220	1111	935	810
Liraglutide	692	330	333	292	285	229	189	192	158	152	121
Sitagliptin	3004	1164	1603	1372	1211	1069	922	829	764	662	580

B