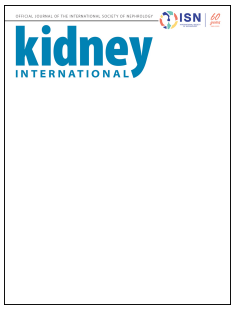


Heterogeneous Treatment Effects of GLP-1RA on Kidney Outcomes in Patients with Type 2 Diabetes

Yang Xu, PhD, Carolien C.H.M. Maas, PhD, Tao Huang, MSc, Tiansheng Wang, PharmD, PhD, Antoine Créon, Arvid Sjölander, Biostat, PhD, Edouard L. Fu, PhD, Juan-Jesus Carrero, Pharm, PhD

PII: S0085-2538(26)00612-5

DOI: https://doi.org/10.1016/j.kint.2026.06.041

Reference: KINT 4784

To appear in: *Kidney International*

Received Date: 26 January 2026

Revised Date: 10 June 2026

Accepted Date: 29 June 2026

Please cite this article as: Xu Y, Maas CCHM, Huang T, Wang T, Créon A, Sjölander A, Fu EL, Carrero JJ, Heterogeneous Treatment Effects of GLP-1RA on Kidney Outcomes in Patients with Type 2 Diabetes, *Kidney International* (2026), doi: https://doi.org/10.1016/j.kint.2026.06.041.

Heterogeneous Treatment Effects of GLP-1RA on Kidney Outcomes in Patients with Type 2 Diabetes

kidney INTERNATIONAL

ISN INTERNATIONAL SOCIETY OF NEPHROLOGY

Study design

Target trial emulation using a new-user, active-comparator design to compare **GLP-1RA** with **DPP-4i**

Population and Outcome

Adults (≥18 years) with type 2 diabetes free from kidney failure

Outcome:
A composite kidney endpoint (sustained ≥40% decline in eGFR, kidney failure, or death due to CKD)

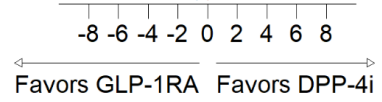
Number of participants:
28,547
(16,775 DPP-4i and 11,772 GLP-1RA)

Method:
Heterogeneous treatment effects through risk modelling with the Kidney Failure Risk Equation (KFRE) and effect modelling with causal survival forest.

Results

The kidney protective effect of GLP-1RA varies by baseline risks

Risk/effect score quarter	GLP-1RA (Event/N)	DPP-4i (Event/N)	Weighted 5-year RD, % (95% CI)	Weighted HR (95% CI)
Overall	306/11772	1125/16775	-0.72 (-1.53 to 0.13)	0.84 (0.72-0.97)
Risk model: KFRE				
Q1 (lowest)	24/3627	58/3510	0.20 (-0.58 to 1.04)	1.20 (0.63-1.94)
Q2	45/3112	143/4025	-0.29 (-1.14 to 0.61)	0.79 (0.54-1.16)
Q3	84/2779	281/4357	0.38 (-1.34 to 2.26)	1.06 (0.77-1.41)
Q4 (highest)	153/2254	643/4883	-2.76 (-5.60 to -0.08)	0.70 (0.56-0.86)
Effect model: CSF				
Q1 (lowest)	18/2848	111/4289	-0.23 (-0.97 to 0.65)	1.06 (0.54-1.74)
Q2	25/3276	118/3860	-0.39 (-1.27 to 0.50)	0.90 (0.50-1.35)
Q3	66/3221	234/3916	-0.91 (-2.07 to 0.47)	0.70 (0.48-0.96)
Q4 (highest)	197/2427	662/4710	-1.09 (-3.85 to 1.94)	0.79 (0.65-0.94)



Abbreviations: ATE, average treatment effect; CATE, conditional average treatment effect; CSF, causal survival forest; DPP-4i, dipeptidyl peptidase-4 inhibitor; GLP-1RA, glucagon-like peptide-1 receptor agonist; HR, hazard ratio; KFRE, 4-Variable non-North America Kidney Failure Risk Equation; RD, risk difference.

Key predictors of treatment benefit

- **Lower** baseline eGFR
- **Faster** preceding eGFR decline
- **Higher** albuminuria
- **Higher** HbA1c
- **Greater** healthcare utilization

Xu et al, 2026

CONCLUSION: *GLP-1RA therapy confers heterogeneous kidney benefits in type 2 diabetes, with high-risk subgroups experiencing the largest risk reduction. Our findings support personalized GLP-1RA prescribing guided by kidney function/damage, trajectories of eGFR decline, and KFRE risk stratification.*

[QUERY TO AUTHOR: title and abstract rewritten by Editorial Office – not subject to change]

Heterogeneous Treatment Effects of GLP-1RA on Kidney Outcomes in Patients with Type 2 Diabetes

Yang Xu, PhD[1,2], Carolien C.H.M. Maas, PhD[3], Tao Huang MSc[1], Tiansheng Wang, PharmD, PhD[4,5], Antoine Créon[2], Arvid Sjölander, Biostat, PhD[2], Edouard L. Fu, PhD[3,2], Juan-Jesus Carrero, Pharm, PhD[2,6]

[1] Department of Pharmacy Administration and Clinical Pharmacy, School of Pharmaceutical Sciences, Peking University, Beijing, China.

[2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

[3] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.

[4] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, USA

[5] Department of Pharmaceutical Health Outcomes and Policy, College of Pharmacy, University of Houston, Houston, USA

[6] Division of Nephrology, Department of Clinical Sciences, Karolinska Institutet, Danderyd Hospital, Stockholm, Sweden.

Short running title: Heterogeneous Kidney Effects of GLP-1RA

Abstract (250-300): 267

Lay Summary (max 150): 150

Main Text Word Count (max 4000): 3994

Number of Tables/Figures (no limit): 5

Reference (no limit): 50

Corresponding author

Yang Xu

Department of Pharmacy Administration and Clinical Pharmacy, School of Pharmaceutical Sciences, Peking University Health Science Center, No. 38 Xueyuan Road, Beijing 100191, China.

Email: xuyang_pucri@bjmu.edu.cn

Telephone: 86-18519121906

Abstract

Introduction: On average, glucagon-like peptide-1 receptor agonists (GLP-1RA) lower the risk of kidney outcomes in type 2 diabetes, but individual responses may differ. Here, we aimed to identify patient subgroups more likely to derive kidney benefit from GLP-1RA therapy.

Methods: Using a target trial emulation framework, we conducted a new-user, active comparator cohort study that included over 28,000 individuals with type 2 diabetes initiating either GLP-1RA or dipeptidyl peptidase-4 inhibitors (DPP-4i) in Stockholm (2008-2021). Heterogeneous treatment effects on a composite kidney outcome were estimated using complementary approaches: risk modelling with the kidney failure risk equation (KFRE) and effect modelling with causal survival forests. For interpretability, Shapley Additive Explanations identified clinical features driving benefit.

Results: We included 11,772 GLP-1RA and 16,775 DPP-4i initiators. Compared with DPP-4i, GLP-1RA initiation was associated with a lower five-year risk of kidney outcomes (risk difference -0.72% [95% confidence interval -1.53 to 0.13]; hazard ratio [HR] 0.84 [0.72-0.97]). Risk modelling revealed pronounced heterogeneity: patients in the highest KFRE-risk quarter experienced a substantial absolute risk reduction (-2.76% [-5.60 to -0.08]; HR 0.70 [0.56-0.86]), whereas those in the lowest quarter showed no benefit. Effect modelling similarly contrasted quarters 4 vs. 1 (HR 0.79 [0.65-0.94] vs. 1.06 [0.54-1.74]). Key predictors of treatment benefit included lower baseline estimated glomerular filtration rate (eGFR), faster preceding eGFR decline, higher albuminuria, higher glycated hemoglobin, and greater healthcare utilization prior to treatment start.

Conclusions: GLP-1RA therapy confers heterogeneous kidney benefits in type 2 diabetes, with certain high-risk subgroups experiencing the largest risk reduction. Such findings support personalized GLP-1RA prescribing guided by kidney function/damage, trajectories of eGFR decline, and KFRE risk stratification.

Keywords: glucagon-like peptide-1 receptor agonists, kidney, heterogeneous treatment effect

Lay summary

People with type 2 diabetes are at increased risk of kidney disease. GLP-1 receptor agonists are known to reduce this risk on average, but not all patients benefit equally. We studied more than 28,000 people with type 2 diabetes in Stockholm who started either a GLP-1 receptor agonist or another common diabetes drug. We followed them for up to five years to see who experienced the largest benefit in terms of kidney-protection. Overall, people taking GLP-1 receptor agonists had a lower risk of kidney complications. However, patients who already had poorer kidney function or higher risk of kidney failure gained the most protection, while those at low risk showed little or no benefit. Key factors linked to greater benefit included worse baseline kidney function, faster recent kidney decline, higher urine protein levels, and more frequent healthcare use before treatment. These findings support a more personalized approach to prescribing these medications.

Main Text

Introduction

Chronic Kidney Disease (CKD) is common among patients with type 2 diabetes, affecting over 25% of individuals,[1] and being associated with various adverse outcomes such as kidney failure with replacement therapy, atherosclerotic cardiovascular disease, heart failure, and mortality.[2] Despite advances in glycemic control and blood pressure management, residual risk for kidney complications remains high, underscoring the need for additional therapies that confer kidney protection beyond glucose lowering.

Glucagon-like peptide-1 receptor agonists (GLP-1RA) have emerged as a cornerstone in the management of type 2 diabetes. In addition to improving glycemic control and reducing cardiovascular events, GLP-1RA exert direct kidney-protective effects by mitigating inflammation, oxidative stress, and fibrosis.[3-5] Clinical trials,[6] most prominently the FLOW trial,[7] as well as observational studies[8, 9] have consistently shown that, on average, GLP-1RA slow estimated glomerular filtration rate (eGFR) decline and reduce risks of macroalbuminuria and kidney failure. However, whether the kidney-protective effects of GLP-1RA extend uniformly across diverse patient populations remains uncertain.

Heterogeneity of treatment effect (HTE) is increasingly recognized in diabetes care. Patients differ in baseline kidney function, albuminuria, glycemic control, and comorbidity burden, all of which may influence the magnitude of benefit derived from GLP-1RA therapy.[10, 11]

Identifying subgroups likely to experience kidney benefit is critical for advancing precision medicine and guiding clinical decision-making. Yet traditional subgroup analyses are ill-suited to capture such multidimensional variation, motivating the use of modern predictive approaches that integrate multiple clinical features to uncover clinically meaningful heterogeneity.[11, 12]

To inform personalized prescribing strategies and refine risk-based approaches to kidney protection in type 2 diabetes, we conducted an observational study comparing GLP-1RA with dipeptidyl peptidase-4 inhibitors (DPP-4i).[13] Using complementary risk and effect modelling approaches, we characterized HTEs on kidney outcomes and identified patient subgroups benefitting from GLP-1RA therapy.

Methods

This study adheres to the Predictive Approaches to Treatment Effect Heterogeneity (PATH) statement for conducting and reporting HTE studies,[11] and follows the target trial emulation framework.[14] The study was approved by the National ethical review board (2017/793-31) and the Swedish National Board of Welfare. The need for informed consent was waived due to the use of de-identified data.

Data source

We used data from the Stockholm CREAtinine Measurements (SCREAM) project, a

comprehensive healthcare utilization cohort of all residents in Stockholm during 2006-2021 (clinicaltrials.gov NCT06239129).[15] The region of Stockholm had a population of 2.3 million citizens in 2021 and provides universal healthcare with a single unified health-system. SCREAM integrates administrative data on demographics, inpatient and outpatient healthcare use, diagnoses, procedures, and vital status. These were enriched with laboratory tests and filled prescriptions at Swedish pharmacies. Registries were linked and de-identified by the Swedish National Board of Welfare with negligible loss to follow-up.

Target trial emulation

We emulated a randomized trial[14] evaluating the effects of GLP-1RA *vs.* DPP-4i on a composite kidney endpoint among patients with type 2 diabetes. DPP-4i were selected as a neutral comparator, as prior trials suggest minimal kidney effects.[14] Key elements of the emulated target trial and the mapping of observational data to trial components are detailed below and in **Supplementary Table S1**.

Study population

We included adults (≥18 years) with type 2 diabetes initiating GLP-1RA or DPP-4i from January 1, 2008 to December 31, 2021. New use was defined as a filled prescription for either a GLP-1RA or a DPP-4i during the study period, with no filled prescription of either drug in the preceding year, as ascertained via the Prescribed Drug Registry.[16] We excluded patients with kidney failure (undergoing dialysis, kidney transplantation or with eGFR <15

ml/min per 1.73 m^2). To capture prior kidney function trajectory as a potential predictor of treatment response, we required ≥2 eGFR measurements in the two years preceding treatment initiation (**Supplementary Table S2**, **Supplementary Figures S1-S2**).

Treatment assignment

Patients who newly filled a prescription for GLP-1RA or DPP-4i were assigned to the corresponding treatment arm. To emulate randomization we adjusted for a number of baseline confounders. Potential confounders were identified via literature review[17, 18] and integrated into a directed acyclic graph (DAG, **Supplementary Figure S3**) to inform the modelling strategy.[19] Baseline confounders included demographics (age, sex, and calendar year), diabetes severity and complications (e.g., duration, neuropathy, and retinopathy identified by ICD-10 codes), comorbid conditions (e.g., hypertension, ischemic heart disease, and heart failure identified by ICD-10 codes), medications (e.g., angiotensin converting enzyme inhibitor/angiotensin receptor blocker, calcium channel blocker, mineralocorticoid receptor antagonist [MRA], diuretic, and sodium-glucose cotransporter-2 inhibitor identified by ATC codes), laboratory tests (glycated hemoglobin [HbA1c], eGFR, urinary albumin-to-creatinine ratio [UACR]) and healthcare utilization (hospitalization, outpatient consultation/visits, and number of drugs dispensed in the previous year). Full list of confounders and detailed operational definitions are provided in **Supplementary Table S3**.

Outcome

The primary outcome was a composite kidney endpoint of sustained ≥40% decline in eGFR, kidney failure (initiation of maintenance dialysis or kidney transplantation or a sustained eGFR <15 ml/min per 1.73 m^2), or death due to CKD (i.e., CKD reported as the main cause of death). To minimize sensitivity to variations in follow-up duration and transient fluctuations in eGFR, we applied a linear interpolation method to ascertain the moment a sustained ≥40% decline in eGFR or persistent eGFR <15 ml/min per 1.73 m^2 was reached.[20] Detailed definitions are provided in **Supplementary Table S4**.

Start and end of follow-up

Follow-up began on the date of first filled prescription (index date, T_0) and continued until the earliest occurrence of the study outcome, death, emigration, or administrative end of follow-up (December 31, 2021).

Causal estimand

The causal estimand was the observational analogue of an intention-to-treat effect (ITT), which was the effect of starting GLP-1RA *vs.* DPP-4i regardless of subsequent treatment discontinuation or switching. Therefore, treatment assignment was based on initiation and did not require persistent use during follow-up. This is similar to the treatment policy strategy in the ICH E9 (R1) estimand framework.[21]

Statistical analysis

We estimated treatment effects at three levels:

1. **Average treatment effect (ATE)**: the overall 5-year risk difference (RD) if all patients received GLP-1RA *vs.* DPP-4i.
2. **Conditional average treatment effect (CATE)**: the treatment effect within clinically relevant subgroups.
3. **Individualized treatment effect (ITE)**: the predicted difference in outcome risk for each patient under GLP-1RA *vs.* DPP-4i.

HTEs were quantified by deviation of CATE from ATE across subgroups and the distribution of ITEs across the cohort (**Supplementary Figure S4**).

Estimation of average treatment effect

To minimize confounding by indication, inverse probability of treatment weighting (IPTW) was used. The probability of receiving GLP-1RA *vs.* DPP-4i was estimated using logistic regression, adjusting for a wide range of 41 confounders (**Supplementary Figure S3**). Stabilized weights were calculated from the propensity scores, and balance was assessed using standardized mean differences (SMD), with an SMD <0.1 indicating negligible imbalance. The 5-year RDs were estimated using weighted Aalen-Johansen estimators, accounting for death as a competing risk, and cause-specific HRs were derived from

weighted Cox proportional hazard models. Confidence intervals (CIs) were obtained via 500 non-parametric bootstrap resamples.

Quantification of heterogenous treatment effects

We assessed HTE using both risk and effect modelling approaches[11] (**Figure 1**). In the risk modelling approach, patients were grouped by baseline kidney failure risk, and treatment effects were then examined across these risk strata. By contrast, in the effect modelling approach, treatment effects were directly estimated using a causal survival forest (CSF).[12]

Risk model through Kidney Failure Risk Equation

Baseline 5-year kidney failure risk was estimated using the externally validated Kidney Failure Risk Equation (KFRE)[22] for non-North American populations, which incorporates age, sex, eGFR, and UACR. Patients were stratified into quarters based on their KFRE risk scores, and IPTW was applied within each quarter to ensure confounder balance. CATEs (both hazard ratios [HR] and 5-year RD) were estimated in each risk quarter. Finally, we estimated ITEs as 5-year RD for each patient by fitting a cause-specific Cox model with treatment, predicted risk score, their interaction term and all confounders identified from the DAG on the original cohort. We first used the fitted model to estimate individualized HRs, and then the corresponding 5-year RDs were calculated using individualized HRs combined with the baseline cumulative hazard derived from Breslow estimator.[23] A distribution of ITEs

was generated to visualize the HTEs of GLP-1RA *vs.* DPP-4i.

Effect model through causal survival forest

CSF, an extension of causal forests for survival data, was used to directly estimate ITEs.[12] Specifically, we constructed 2000 causal survival trees, incorporating treatment, 41 potential confounders, and 2 additional candidate predictors that captured prior kidney function trajectory (**Supplementary Table S3**) as independent variables, with the composite kidney outcome as the dependent variable. For each tree, we randomly drew 50% of the cohort without replacement and further split this subsample into two equal parts. The first part was used to determine the tree structure by maximizing treatment effect heterogeneity, while the second part was used to estimate leaf-specific treatment effects. ITE estimates were obtained by averaging leaf-level estimates across trees. Key hyperparameters, the number of variables considered for each split and the minimum number of observations in each terminal leaf node, were tuned via a grid search and evaluated using out-of-bag (OOB) validation. Implementation details are provided in **Supplementary Methods S1**, and elsewhere.[24] The CSF model generated individualized 5-year RD (i.e., ITE) as the effect score for each patient. Patients were ranked by these effect scores, stratified into quarters, and then CATEs were estimated. Deviations from the ATE for each ITE quarter and the histogram of ITEs were used to quantify the HTEs of GLP-1RA *vs.* DPP-4i.

Model evaluation

Risk and effect models were evaluated for discrimination (the ability to discriminate between patients with small *vs.* large treatment effects, measured by C-for-benefit, where higher values indicate better performance) and calibration (the agreement between observed and predicted treatment effects, measured by E_{50}-for-benefit, where lower values indicate better agreement),[25] supported by visual calibration-for-benefit plots.[25] Internal validation was performed using 500 bootstrap replicates, with 95% CIs derived from the bootstrap distributions (**Supplementary Methods S2**).[26]

Model interpretation

To enhance the clinical interpretability, we employed two complementary methods. First, we compared baseline characteristics of patients across quarters of risk scores and effect scores, respectively. Second, we assessed the variable importance for each predictor on ITE predictions of the effect model using the model-agnostic Shapley Additive Explanations (SHAP) method.[27] SHAP quantifies each predictor's importance by calculating a local Shapley value per predictor and patient. This value estimates how much the ITE changes when a predictor is perturbed from its observed value to the sample mean, averaging over all possible combinations of other predictors. A higher absolute value of a local Shapley value means greater contribution of the patient's predictor value to its ITE prediction, with the sign describing whether that patient's predictor value is accountable for making the ITE higher (increased benefit from DPP-4is) or lower (increased benefit from GLP-1RAs). Besides, for

each predictor, the mean absolute local Shapley values across all patients (global Shapley value) measures the predictor's global importance on ITE prediction, with higher values suggest greater global importance. Detailed implementation steps are provided in **Supplementary Methods S3**.

Sensitivity analyses

We conducted quantitative bias analyses to assess how strong unmeasured confounding would need to be to explain away the observed ATE and CATE estimates across risk and effect quartiles. Specifically, we introduced a hypothetical binary unmeasured confounder (prevalence fixed at 0.5) and reestimated the treatment effects. By varying its associations with both treatment and the outcome, we evaluated the strength of confounding required to attenuate the observed 5-year RD to the null.[28] Detailed methods are provided in **Supplementary Methods S4**.

All analyses were conducted using R (version 4.3.2, R Foundation for Statistical Computing, Vienna, Austria).

Results

Baseline characteristics

A total of 28,547 adults with type 2 diabetes initiating treatment between 2008 and 2021 were included, comprising 16,775 DPP-4i initiators and 11,772 GLP-1RA initiators

(**Supplementary Figure S2**). Compared with DPP-4i initiators, GLP-1RA initiators were younger (61 *vs.* 67 years), had higher rates of hypertension (62% *vs.* 56%) and obesity (27% *vs.* 8%), higher HbA1c (61 *vs.* 58 mmol/mol) and eGFR (91 *vs.* 85 mL/min per 1.73 m^2), while their preceding kidney function decline trajectories were comparable (eGFR decline 0.82 *vs.* 1.02 mL/min per 1.73 m^2 per year; log UACR decline 0.01 *vs.* 0.00 mg/g/year). After IPTW, all baseline confounders were well-balanced between GLP-1RA initiators and DPP-4i initiators, with SMDs <0.10 (**Table 1,**[29] **Supplementary Table S5,** and **Supplementary Figure S5**). Propensity score distributions showed sufficient overlap without excessively large weights (**Supplementary Figure S6**).

Average treatment effect

During a median (interquartile range [IQR]) follow-up of 2.2 (0.9-4.4, for GLP-1RA initiators) and 4.2 (2.2-7.1, for DPP-4i initiators) years, 306 kidney events occurred among 11,772 GLP-1RAs initiators and 1,125 kidney events occurred among 16,775 DPP-4i initiators, corresponding to incidence rates of 8.4 (95% CI 7.5-9.3) and 13.6 (12.8-14.4) per 1000 person-years, respectively. Compared with DPP-4i, GLP-1RA use was associated with a 0.72% absolute reduction in the risk of the composite kidney outcome at 5 years (weighted 5-year RD -0.72% [-1.53% to 0.13%]), translating to a weighted HR of 0.84 (0.72-0.97) (**Table 2**, **Supplementary Figure S7**).

Heterogeneous treatment effects based on risk modelling

KFRE-predicted risk scores exhibited similar distributions between treatment arms, with most patients clustered in low-risk categories (**Supplementary Figure S8**). Patients in the highest-risk quarter (Q4) were older, had higher prevalence of hypertension, more severe diabetes, and worse kidney function compared with those in the lowest-risk quarter (Q1) (**Supplementary Table S6**). After IPTW, all covariates were well-balanced (SMDs <0.10) with adequate propensity score overlap and no extreme weights (**Supplementary Figures S9-S10**).

We observed important treatment heterogeneity, with a larger reduction in the risk of composite kidney outcomes amongst individuals with highest KFRE-predicted risk scores (**Table 2**, **Supplementary Figures S11-S12**). Specifically, the weighted 5-year RD in Q4 was -2.76% (95% CI -5.60 to -0.08) and HR 0.70 (0.56-0.86), corresponding to a deviation from the ATE of -2.04% (-4.21 to 0.07). In contrast, lower-risk quarters (Q1-Q3) demonstrated minimal or nonsignificant treatment effects (weighted 5-year RD 0.20% [-0.58 to 1.04] in Q1, -0.29% [-1.14 to 0.61] in Q2, 0.38% [-1.34 to 2.26] in Q3) (**Table 2**, **Supplementary Figures S11-S12**). The distribution of ITEs showed increasing treatment benefit with higher baseline KFRE risk (**Figure 2a**).

Heterogeneous treatment effects based on effect modelling

Stratification by CSF-predicted effect scores revealed patterns consistent with KFRE risk

strata: Q4 patients were older, had higher prevalence of hypertension, more severe diabetes, and lower kidney function (**Supplementary Table S7**). Following IPTW, all covariates were well-balanced (SMDs <0.10) with adequate propensity score overlap and reasonable weights (**Supplementary Figures S13-S14**). Effect modelling identified enhanced benefit in Q3-Q4. Weighted 5-year RDs were -0.91% (95% CI -2.07% to 0.47%) in Q3 and -1.09% (-3.85% to 1.94%) in Q4, with significantly lower HRs of 0.70 (0.48-0.96) and 0.79 (0.65-0.94), respectively. No significant treatment effect was observed in Q1-Q2 (**Table 2, Supplementary Figures S11 & S15**).

Model evaluation and interpretation

The CSF model showed better discrimination than the KFRE-based risk model (C-for-benefit, 0.605 [95% CI 0.464-0.614] *vs.* 0.532 [0.466-0.599]), while maintaining comparable calibration (E_{50}-for-benefit 0.009 [0.004-0.015] *vs.* 0.010 [0.004-0.017], **Table 2** and **Supplementary Figure S16**). Shapley value analysis for the effect model CSF identified lower baseline eGFR, faster preceding eGFR decline, higher albuminuria, worse glycemic control (higher HbA1c) and more frequent healthcare use (collectively a higher number of medications dispensed, outpatient or hospital visits in the year prior to treatment start) as the strongest predictors of benefit from GLP-1RA (**Figure 3**).

Sensitivity analyses

In quantitative bias analyses, all observed combinations of sensitivity parameters for

measured confounders laid within the red boundaries of the contour plot (**Supplementary Figure S17**), indicating that unmeasured confounding sufficient to nullify the observed ATE would need to be substantially stronger than measured confounding. Similar patterns were observed for the CATE across KFRE risk quarters and CSF effect quarters (**Supplementary Figures S18-S19**).

Discussion

In this large cohort of patients with type 2 diabetes treated by routine care in Sweden, we quantified the HTEs of GLP-1RA *vs.* DPP-4i on reducing kidney risks. Using two complementary analytical approaches, we found that patients with the baseline highest risk derived the greatest kidney benefit from GLP-1RA. Specifically, treatment benefit was most pronounced among individuals with lower eGFR, higher albuminuria, faster eGFR decline and worse glycemic control. Conversely, healthier/lower-risk individuals did not benefit as much from use of GLP-1RA. Our findings highlight the value of moving beyond a uniform treatment strategy toward a data-informed, individualized approach to optimize kidney protection.

Our target trial emulation benchmarked the average kidney-protective effect of GLP-1RA in patients with type 2 diabetes, with a HR of 0.84 that aligns closely with prior meta-analyses of randomized trials.[6] We expand this evidence by further exploring the heterogeneity of treatment benefit. Traditional subgroup analyses seldom identify meaningful HTEs, because

they evaluate one variable at a time, neglect interactions among patient characteristics, and often lack sufficient statistical power, increasing the risk of false-negative findings.[30, 31] Furthermore, whereas the FLOW trial was the sole dedicated trial on GLP-1RAs and kidney outcomes and included patients with impaired kidney function,[7] previous trials primarily enrolled diabetic patients with high cardiovascular and preserved eGFR and CKD outcomes were secondary/post-hoc.[6] Thus, the homogeneity of participants included in existing trials may have further limited their ability to detect HTE.

Using classic approaches (looking at one condition at a time through subgroup analyses), there is some preceding evidence of heterogeneity of effect of GLP-1RA: GLP-1RA use has been associated with higher HbA1c reduction and lower hypoglycemia risk in older *vs.* younger patients[32, 33]; weight losses attributed to GLP-1RA have been reported larger in female *vs.* male patients[34]; and cardioprotection from GLP-1RA use has been observed to be larger in people with higher body mass index[35] or on those with atherosclerotic cardiovascular disease.[36] Our study expands this evidence by identifying and ranking predictors of kidney benefit. Our first finding was that patients with albuminuria and moderate-to-severe CKD experienced the greatest benefit, which agrees with subgroup analyses of trials.[37, 38]

Our second and novel finding was that patients with a preceding rapid decline in eGFR featured as deriving highest treatment benefit. Conceptually, such observation alignedwith a

post-hoc analysis of the SONAR trial, where the efficacy of atrasentan, a selecting endothelin A receptor antagonist, in slowing progressive kidney function loss was larger in participants with a steeper pretrial eGFR decline.[39] Also in an observational study of patients with IgA nephropathy, the benefit of dapagliflozin, a sodium-glucose cotransporter 2 inhibitor, in delaying eGFR decline was determined by the pre-treatment eGFR slope.[40] Although these findings may seem intuitive, further work is needed to evaluate whether preceding eGFR slope could be a valid criterion for enrolment in clinical trials or to help guide individualized therapy. As suggested in the PERL trial,[41] a rapid decline prior to study entry may reflect regression to the mean or behavioral changes before randomization, which can attenuate the subsequent eGFR decline and complicate interpretation.

We observed that participants with poor glycemic control also experienced larger treatment benefit, consistent with the ACCORD trial, which observed that individuals with the highest baseline KFRE risk best benefited from intensive glycemic control and showed the largest reductions in the rate of kidney, microvascular outcomes and all-cause mortality.[42] Finally, sicker participants (those with more frequent healthcare interactions or more medications) experienced larger treatment benefits in our study. In line with this, the potential benefit of GLP-1RAs on reducing asthma exacerbation relative to sulfonylureas was reported to be larger amongst more severe cases (with more frequent history of emergency contacts) and those with the highest number of prescriptions for short-acting β_2-agonists or inhalers.[43]

Although CSF identified several effect modifiers not included in the KFRE that were consistent with prior evidence, findings from effect modelling approaches, particularly machine learning–based methods such as CSF, should be interpreted with caution due to their susceptibility to overfitting, particularly in settings with relatively limited numbers of outcome events. Prior studies have shown that findings on HTE from effect models can vary depending on the choice of algorithm or model specification.[44, 45] In contrast, risk modelling approach may provide more robust and clinically interpretable assessments of HTE because they are generally simpler and supported by stronger clinical and theoretical rationale.[46] Nevertheless, both approaches yield consistent findings in our study suggesting that patients with higher baseline kidney risk may be the ones deriving greater kidney benefit from GLP-1RA treatment. We conclude that while risk modelling should be prioritized for reliable HTE detection, we interpret that effect modelling is more useful for hypothesis generation. Before informing clinical practice, the predictors of GLP-1RA benefit identified in our CSF model should thus undergo replication in future studies.

This study has several strengths: We leveraged rich healthcare data from a complete geographical region with universal care, which may be less susceptible to data fragmentation and socioeconomic disparities in healthcare access; We implemented a rigorous causal inference framework including a target trial emulation with an active comparator, new-user design to mitigate time-related and selection biases, and rich clinical data including a battery of laboratory tests that were utilized for confounding control and HTE assessments; We

applied a state-of-the-art CSF, which accounts for selection biases during follow-up while assessing HTEs in time-to-event data, thereby reducing the impact of informative censoring.

Nonetheless, there are aspects that desereve consideration when interpreting our findings: First, residual confounding cannot be ruled out from any observational study. However, quantitative bias analyses suggested that substantial unmeasured confounding would likely be required to fully explain away the observed results. HTE estimation in observational data remains dependent on the assumption of no unmeasured confounding, and no established method currently exists to quantify violations of this assumption in ITE estimation. Such bias may create spurious heterogeneity or obscure true heterogeneity.[47] Moreover, despite the inclusion of 43 clinically relevant predictors in the CSF model, residual heterogeneity may remain if important effect modifiers are unobserved—for example, genetic predisposition, physical activity, or ethnicity. In such cases, true HTE could go undetected (false negatives). Second, because kidney events are long-term and take years to develop, our analysis followed an ITT principle and did not explicitly account for treatment discontinuation, switching, or augmentation. This may bias the effect estimates toward the null,[48] but current statistical methods do not yet allow for censoring due to treatment deviations[12]; When estimating ITEs, current methods do neither allow for simultaneous adjustment for the competing risk of death, which may lead to overestimation of both the absolute benefit of GLP-1RAs (given more non-CKD-related deaths in the DPP-4i group) and the proportion of patients who benefit.[49] Third, although the KFRE was originally developed for patients with eGFR <60

mL/min per 1.73 m², we did not find other validated risk equation tools that could be applied in our data. Furthermore, recent evidence suggests that KFRE maintains good discrimination (C-statistics 0.75) in people eGFR ≥60 mL/min per 1.73 m².[50] Therefore, because the KFRE was used in our study primarily for risk stratification rather than absolute risk prediction, our findings remain informative as long as the relative ordering of kidney failure risk is reasonably preserved. Finally, the external validity of our predictive models and the biological underpinnings of the observed heterogeneity warrant exploration in future studies before they can inform clinical practice. Although we focused on one aspect of GLP-1RA therapy, the overall benefit–risk balance should warrant integrated consideration of cardiovascular and kidney outcomes, alongside clinically significant adverse effects such gastrointestinal intolerance and pancreatitis. In our cohort, gastrointestinal intolerance was not reliably captured by ICD-10 codes, and pancreatitis was rare, which limited our capacity to explore heterogeneity across adverse event risks. These limitations motivate two key next steps: external validation of our findings across diverse populations to enhance generalizability, and if confirmed, conduct of prospective trials to test the hypothesis that prioritizing GLP-1RA to high-risk patients maximizes health benefits, minimizes risks and is cost-effective.

To conclude, this study demonstrates that the kidney-protective effect of GLP-1RA in patients with type 2 diabetes is not uniform, exhibiting substantial heterogeneity. We identified that patients with high baseline kidney risk, particularly those with low eGFR,

rapid preceding eGFR decline, and elevated albuminuria, derived most benefits from GLP-1RA therapy. Prescribing treatments based on risks may better align with the principles of personalized medicine and supports efficient resource allocation by targeting therapies toward patients with the lowest number needed to treat.

Disclosure statement

None of the other authors declares relevant financial or scientific interests that would represent a conflict of interest.

Data Availability Statement

The data underlying this article cannot be shared publicly due to the privacy of individuals that participated in the study. The data may be shared on reasonable request for academic research collaborations that fulfil GDPR, as well as national and institutional ethics regulations and standards by contacting Prof. Juan-Jesus Carrero (juan.jesus.carrero@ki.se).

Acknowledgments

The authors thank all staff for their contributions and the study participants for their participation.

Funding Information

Y.X. is supported by the Young Scientists Fund grant from National Natural Science Foundation of China (Grant no. 82304245), and the Key Laboratory of Epidemiology of Major Diseases (Peking University) grant from Ministry of Education of China.

E.L.F. is supported by a VENI grant (09150162310058), a Junior Kolff grant (22OK2026) from the Dutch Kidney Foundation and a Junior Principal Investigator

grant from Leiden University Medical Center.

T.W. is supported by American Diabetes Association grant #4-22-PDFPM-06 and a grant from North Carolina Diabetes Research Center (P30DK124723).

J.J.C. and the SCREAM project are supported by the Swedish Research Council, the Swedish Heart and Lung Foundation (#20250416), and Region Stockholm.

The funders had no role in the design or conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author Contributions

Y.X. had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Y.X., E.L.F., J.J.C..

Acquisition of data: J.J.C..

Analysis and interpretation of data: all authors.

Drafting of the manuscript: Y.X., J.J.C., C.C.M., T.W., T.H., A.C..

Critical revision of the manuscript for important intellectual content: all authors.

Statistical analysis: Y.X., T.H., A.S..

Administrative, technical or material support: all authors.

Supervision: Y.X., J.J.C..

Supplementary material is available online at www.kidney-international.org.

References

1. de Boer IH, Khunti K, Sadusky T*, et al.* Diabetes management in chronic kidney disease: a consensus report by the American Diabetes Association (ADA) and Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int* 2022; **102:** 974-989.

2. Fox CS, Matsushita K, Woodward M*, et al.* Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without diabetes: a meta-analysis. *Lancet* 2012; **380:** 1662-1673.

3. Mann JFE, Muskiet MHA. Incretin-based drugs and the kidney in type 2 diabetes: choosing between DPP-4 inhibitors and GLP-1 receptor agonists. *Kidney International* 2021; **99:** 314-318.

4. Tuttle KR, Wilson JM, Lin Y*, et al.* Indicators of Kidney Fibrosis in Patients with Type 2 Diabetes and Chronic Kidney Disease Treated with Dulaglutide. *Am J Nephrol* 2023; **54:** 74-82.

5. Alicic RZ, Neumiller JJ, Tuttle KR. Mechanisms and clinical applications of incretin therapies for diabetes and chronic kidney disease. *Curr Opin Nephrol Hypertens* 2023; **32:** 377-385.

6. Badve SV, Bilal A, Lee MMY*, et al.* Effects of GLP-1 receptor agonists on kidney and cardiovascular disease outcomes: a meta-analysis of randomised controlled trials. *Lancet Diabetes Endocrinol* 2025; **13:** 15-28.

7. Perkovic V, Tuttle KR, Rossing P*, et al.* Effects of Semaglutide on Chronic Kidney Disease in Patients with Type 2 Diabetes. *N Engl J Med* 2024; **391:** 109-121.

8. Xie Y, Bowe B, Gibson AK*, et al.* Comparative Effectiveness of SGLT2 Inhibitors, GLP-1 Receptor Agonists, DPP-4 Inhibitors, and Sulfonylureas on Risk of Kidney Outcomes: Emulation of a Target Trial Using Health Care Databases. *Diabetes Care* 2020; **43:** 2859-2869.

9. Xu Y, Fu EL, Clase CM*, et al.* GLP-1 receptor agonist versus DPP-4 inhibitor and kidney and cardiovascular outcomes in clinical practice in type-2 diabetes. *Kidney International* 2022; **101:** 360-368.

10. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet* 1995; **345:** 1616-1619.

11. Kent DM, van Klaveren D, Paulus JK*, et al.* The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration. *Ann Intern Med* 2020; **172:** W1-w25.

12. Cui Y, Kosorok MR, Sverdrup E*, et al.* Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2023; **85:** 179-211.

13. Gong Y, Bai X, Zhang D*, et al.* Effect of DPP-4i inhibitors on renal function in patients with type 2 diabetes mellitus: a systematic review and meta-analysis of randomized controlled trials. *Lipids in Health and Disease* 2024; **23:** 157.

14. Hernán MA, Dahabreh IJ, Dickerman BA*, et al.* The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful? *Ann Intern Med* 2025; **178:** 402-407.

15. Carrero JJ, Elinder CG. The Stockholm CREAtinine Measurements (SCREAM) project: Fostering impr ovements in chronic kidney disease care. *J Intern Med* **291:** 254-268.

16. Wettermark B, Hammar N, Fored CM*, et al.* The new Swedish Prescribed Drug Register--opportunities for pharmacoepidemiological research and experience from the first six months. *Pharmacoepidemiol Drug Saf* 2007; **16:** 726-735.

17. Bajaj M, McCoy RG, Balapattabi K*, et al.* 11. Chronic Kidney Disease and Risk Management: Standards of Care in Diabetes—2026. *Diabetes Care* 2026.

18. Kazancioğlu R. Risk factors for chronic kidney disease: an update. *Kidney Int Suppl (2011)* 2013; **3:** 368-371.

19. Tennant PWG, Murray EJ, Arnold KF*, et al.* Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol* 2021; **50:** 620-632.

20. Carrero JJ, Fu EL, Vestergaard SV*, et al.* Defining measures of kidney function in observational studies using routine health care data: methodological and reporting considerations. *Kidney Int* 2023; **103:** 53-69.

21. Kahan BC, Hindley J, Edwards M*, et al.* The estimands framework: a primer on the

ICH E9(R1) addendum. *BMJ* 2024; **384:** e076316.

22. Tangri N, Grams ME, Levey AS*, et al.* Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis. *JAMA* 2016; **315:** 164-174.

23. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 1999; **94:** 496-509.

24. Inoue K, Seeman TE, Horwich T*, et al.* Heterogeneity in the association between the presence of coronary artery calcium and cardiovascular events: A machine-learning approach in the MESA study. *Circulation* 2023; **147:** 132-141.

25. Maas CCHM, Kent DM, Hughes MC*, et al.* Performance metrics for models designed to predict treatment effect. *BMC medical research methodology* 2023; **23:** 165.

26. Desai RJ, Glynn RJ, Solomon SD*, et al.* Individualized treatment effect prediction with machine learning — salient considerations. *NEJM Evidence* 2024; **3**.

27. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, Curran Associates Inc., 2017, pp 4768–4777

28. Huang R, Xu R, Dulai PS. Sensitivity analysis of treatment effect to unmeasured confounding in observational studies with survival and competing risks outcomes. *Statistics in medicine* 2020; **39:** 3397-3411.

29. Sumida K, Nadkarni GN, Grams ME*, et al.* Conversion of Urine Protein-Creatinine Ratio or Urine Dipstick Protein to Urine Albumin-Creatinine Ratio for Use in Chronic Kidney Disease Screening and Prognosis : An Individual Participant-Based Meta-analysis. *Ann Intern Med* 2020; **173:** 426-435.

30. Brookes ST, Whitely E, Egger M*, et al.* Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004; **57:** 229-236.

31. Kent DM, Nelson J, Dahabreh IJ*, et al.* Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *International journal of epidemiology* 2016; **45:** 2075-2088.

32. Hanlon P, Butterly E, Wei L*, et al.* Age and sex differences in efficacy of treatments for type 2 diabetes: A network meta-analysis. *JAMA* 2025; **333:** 1062-1073.

33. Dávila-García DM, Falconer T, Pratt N*, et al.* Heterogeneity of treatment effects of glucose-lowering drug classes for type 2 diabetes: LEGEND-T2DM network real-world evidence. *J Diabetes Complications* 2025; **39:** 109114.

34. Marassi M, Cignarella A, Russo GT*, et al.* Sex differences in the weight response to GLP-1RA in people with type 2 diabetes. A long-term longitudinal real-world study. *Pharmacol Res* 2025; **219:** 107866.

35. Tang ASP, Hsu JTY, Chong SKS*, et al.* Glucagon-like peptide-1 receptor agonist in myocardial infarction and atherosclerotic cardiovascular disease risk reduction: a comprehensive meta-analysis of number needed to treat, efficacy and safety. *Cardiovasc Diabetol* 2025; **24:** 285.

36. Neugebauer R, An J, Dombrowski SK*, et al.* Glucose-lowering medication classes and cardiovascular outcomes in patients with type 2 diabetes. *Jama Netw Open* 2025; **8:** e2536100.

37. Shaman AM, Bain SC, Bakris GL*, et al.* Effect of the Glucagon-Like Peptide-1 Receptor Agonists Semaglutide and Liraglutide on Kidney Outcomes in Patients With Type 2 Diabetes: Pooled Analysis of SUSTAIN 6 and LEADER. *Circulation* 2022; **145:** 575-585.

38. Chen J-Y, Hsu T-W, Liu J-H*, et al.* Kidney and cardiovascular outcomes among patients with CKD receiving GLP-1 receptor agonists: A systematic review and meta-analysis of randomized trials. *American Journal of Kidney Diseases* 2025; **85:** 555-569.e551.

39. Waijer SW, de Vries ST, Busch R*, et al.* Large Between-Patient Variability in eGFR Decline before Clinical Trial Enrollment and Impact on Atrasentan's Efficacy: A Post Hoc Analysis from the SONAR Trial. *J Am Soc Nephrol* 2021; **32:** 2731-2734.

40. Ueda H, Tsuboi N, Matsumoto K*, et al.* Pre-treatment GFR slopes predict dapagliflozin response in IgA nephropathy. *Kidney International Reports* 2025.

41. Doria A, Galecki AT, Spino C*, et al.* Serum Urate Lowering with Allopurinol and Kidney Function in Type 1 Diabetes. *N Engl J Med* 2020; **382:** 2493-2503.

42. Charu V, Liang JW, Chertow GM*, et al.* Heterogeneous Treatment Effects of Intensive Glycemic Control on Kidney Microvascular Outcomes and Mortality in ACCORD. *J Am Soc Nephrol* 2024; **35:** 216-228.

43. Wang T, Keil AP, Buse JB*, et al.* Glucagon-like Peptide 1 Receptor Agonists and

Asthma Exacerbations: Wh ich Patients Benefit Most? *Annals ATS* **21:** 1496-1506.

44. Bouvier F, Peyrot E, Balendran A*, et al.* Do machine learning methods lead to similar individualized treatment rules? A comparison study on real data. *Statistics in medicine* 2024; **43:** 2043-2061.

45. Sinha P, Spicer A, Delucchi KL*, et al.* Comparison of machine learning clustering algorithms for detecting heterogeneity of treatment effect in acute respiratory distress syndrome: A secondary analysis of three randomised controlled trials. *EBioMedicine* 2021; **74:** 103697.

46. Selby JV, Maas CCHM, Fireman BH*, et al.* Predictive Modeling of Heterogeneous Treatment Effects in RCTs: A Scoping Review. *JAMA Network Open* 2025; **8:** e2522390-e2522390.

47. Stürmer T, Rothman KJ, Avorn J*, et al.* Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study. *American journal of epidemiology* 2010; **172:** 843-854.

48. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clinical trials (London, England)* 2012; **9:** 48-55.

49. Berry SD, Ngo L, Samelson EJ*, et al.* Competing risk of death: an important consideration in studies of older adults. *Journal of the American Geriatrics Society* 2010; **58:** 783-787.

50. Grams ME, Brunskill NJ, Ballew SH*, et al.* The Kidney Failure Risk Equation: Evaluation of Novel Input Variables including eGFR Estimated Using the CKD-EPI 2021 Equation in 59 Cohorts. *J Am Soc Nephrol* 2023; **34:** 482-494.

Figure legends

Figure 1. Graphical illustration of two modelling strategies (a) Risk model: KFRE and (b) Effect model: CSF for heterogeneous treatment effect quantification.

Abbreviations: CATE, conditional average treatment effect; CSF, causal survival forest; DPP-4i, dipeptidyl peptidase-4 inhibitor; eGFR, estimated glomerular filtration rate; GLP-1RA, glucagon-like peptide-1 receptor agonist; HbA1c, glycated hemoglobin; ITE, individualized treatment effect; KFRE, Kidney Failure Risk Equation; RD, risk difference; UACR, urinary albumin-to-creatinine ratio.

Figure 2. Distribution of ITEs predicted by (a) External risk model: KFRE and (b) Effect model: CSF.

Abbreviations: CSF, causal survival forest; DPP-4i, dipeptidyl peptidase-4 inhibitor; GLP-1RA, glucagon-like peptide-1 receptor agonist; ITE, individualized treatment effect; KFRE, Kidney Failure Risk Equation.

Note:

In the risk model, patients with a higher predicted baseline KFRE risk generally experienced greater treatment benefit (ITE <0; **Panel a**). However, all 207 (0.7%) patients who derived no benefit or experienced harm (ITE ≥0) were also from the highest-risk quarter. Conversely, in the effect model, all 1,309 (4.6%) patients with no benefit or harm were from the lowest-effect quarter (**Panel b**).

Figure 3. Shapley values for predictors included in the effect model of CSF: (a) Relative importance of predictors measured by global Shapley value and (b) Association between predictor value and ITE prediction for each patient measured by local Shapley value.

Abbreviations: ACEi, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; CCB, calcium channel blockers; COPD, chronic obstructive pulmonary disease; CSF, causal survival forest; DPP-4i, dipeptidyl peptidase-4 inhibitor; eGFR, estimated glomerular filtration rate; GLP-1RA, glucagon-like peptide-1 receptor agonist; HbA1c, glycated hemoglobin; ITE, individualized treatment effect as 5-year differences; MRA, mineralocorticoid receptor antagonist; NSAID, nonsteroidal anti-inflammatory drug; SGLT-2i, sodium-glucose cotransporter-2 inhibitor; UACR, urinary albumin-to-creatinine ratio.

Note:

Predictors with higher global Shapley values (mean absolute of the local shapely vales for each predictor) in **Panel a** indicate greater importance in predicting treatment heterogeneity, which corresponds to wider distributions along the x-axis in the beeswarm plot (**Panel b**). In **Panel b**, within each bar of specific predictor, each dot represents a single patient, colored by the predictor's scaled value (range from 0 [blue] to 1 [red]) and positioned along the x-axis by its local Shapley contribution. Local Shapley value <0 for a patient suggests that patient's specific predictor value contributes to an increased predicted treatment benefit, defined as a larger 5-year risk

reduction of composite kidney outcome with GLP-1RA compared to DPP-4i, while value >0 indicates a reduced benefit. Therefore, more red dots with local Shapley values <0 and blue dots with local Shapley values >0 suggest that higher predictor value will increase the predicted treatment benefit (e.g., UACR), while more red dots with local Shapley values >0 and blue dots with local Shapley values <0 suggest that lower predictor value will increase the predicted treatment benefit (e.g., eGFR).

Table

Table 1. Selected baseline characteristics of new users of GLP-1RA and DPP-4i before and after inverse probability of treatment weighting.

	Before IPTW			After IPTW		
Treatment arms	**DPP-4i**	**GLP-1RA**	**SMD**	**DPP-4i**	**GLP-1RA**	**SMD**
Total number of patients	16775	11772		16647	11245	
Demographic						
Age, years, mean (SD)	66.77 (12.33)	61.04 (11.81)	**0.475**	64.40 (13.05)	63.58 (11.31)	0.067
Female, n (%)	6669 (39.8)	4988 (42.4)	0.053	6774 (40.7)	4584 (40.8)	0.002
Diabetes severity						
Diabetes duration, years, median [IQR]	6.75 [2.99, 10.84]	6.54 [2.68, 11.43]	0.022	6.71 [2.85, 11.11]	6.69 [2.96, 11.13]	0.002
Comorbid condition, n (%)						
Hypertension	9314 (55.5)	7298 (62.0)	**0.132**	9692 (58.2)	6603 (58.7)	0.010
Ischemic heart disease	1757 (10.5)	1268 (10.8)	0.010	1754 (10.5)	1151 (10.2)	0.010
Heart failure and cardiomyopathy	1213 (7.2)	704 (6.0)	0.050	1107 (6.7)	712 (6.3)	0.013
Peripheral vascular disease	434 (2.6)	284 (2.4)	0.011	416 (2.5)	260 (2.3)	0.012
Cerebrovascular disease	751 (4.5)	438 (3.7)	0.038	690 (4.1)	452 (4.0)	0.007
Obesity	1281 (7.6)	3201 (27.2)	**0.534**	2450 (14.7)	1870 (16.6)	0.053
Ongoing medication use, n (%)						
ACEi/ARB	11419 (68.1)	8362 (71.0)	0.064	11530 (69.3)	7797 (69.3)	0.002
Diuretic	4037 (24.1)	2564 (21.8)	0.054	3844 (23.1)	2529 (22.5)	0.014
MRA	1025 (6.1)	817 (6.9)	0.034	1065 (6.4)	729 (6.5)	0.003
Beta-blocker	7398 (44.1)	4753 (40.4)	0.075	7128 (42.8)	4723 (42.0)	0.017

Antiplatelet agent	5614 (33.5)	3103 (26.4)	**0.156**	5126(30.8)	3352 (29.8)	0.021
Anticoagulant	2096 (12.5)	1288 (10.9)	0.048	1951 (11.7)	1265 (11.3)	0.015
Lipid lowering drug	10950 (65.3)	7778 (66.1)	0.017	10933(65.7)	7412 (65.9)	0.005
Antidiabetic medication use, n (%)						
Metformin	13104 (78.1)	9379 (79.7)	0.038	13181 (79.2)	8944 (79.5)	0.009
Sulfonylurea	4347 (25.9)	2073 (17.6)	**0.202**	3840 (23.1)	2538 (22.6)	0.012
SGLT-2i	782 (4.7)	1344 (11.4)	**0.250**	1278 (7.7)	898 (8.0)	0.011
Insulin	2449 (14.6)	3837 (32.6)	**0.434**	3772 (22.7)	2642 (23.5)	0.020
Laboratory test, median [IQR]						
HbA1c, mmol/mol	58.2 [51.8, 66.6]	60.5 [51.8, 71.3]	**0.153**	59.1 [52.2, 68.7]	59.9 [52.0, 69.7]	0.015
eGFR, ml/min per 1.73 m^2 [a]	84.6 [65.1, 96.8]	91.2 [76.0,101.6]	**0.336**	87.44 [68.5, 99.8]	88.1 [71.4, 98.9]	0.047
CKD stage, n (%)						
G1	6600 (39.3)	6192 (52.6)		7508 (45.1)	5139 (45.7)	
G2	6791 (40.5)	4230 (35.9)		6193 (37.2)	4520 (40.2)	
G3a	2106 (12.6)	937 (8.0)		1831 (11.0)	1102 (9.8)	
G3b	1116 (6.7)	373 (3.2)		982 (5.9)	439 (3.9)	
G4-5	162 (1.0)	40 (0.3)		133 (0.8)	45 (0.4)	
UACR, mg/g [b]	16.4 [7.4, 34.2]	14.18 [6.9, 32.5]	0.024	15.87 [7.3, 33.6]	14.9 [7.0, 33.2]	0.006
Albuminuria category, n (%)						
A1	9880 (58.9)	7677 (65.2)		12086 (72.6)	8231 (73.2)	
A2	3126 (18.6)	2302 (19.6)		3695 (22.2)	2496 (22.2)	
A3	691 (4.1)	449 (3.8)		866 (5.2)	517 (4.6)	
Missing	3078 (18.3)	1344 (11.4)		0 (0)	0 (0)	
Preceding kidney function trajectory, median [IQR]						
eGFR decline rate, ml/min per 1.73 m^2	1.02 [0.15, 2.27]	0.82 [0.11, 1.76]	0.007	0.95 [0.16, 2.12]	0.88 [0.10, 1.98]	0.005

per year						
log UACR decline rate, mg/g/year	0.00 [-0.15, 0.13]	0.01 [-0.12, 0.14]	0.007	0.00 [-0.14, 0.15]	0.01 [-0.13, 0.15]	0.004
Healthcare utilization						
No. of drugs used, median [IQR]	11 [7, 15]	11 [8, 16]	0.077	11 [8, 16]	11 [8, 15]	0.002
Outpatient consultation, n (%)	5856 (34.9)	4863 (41.3)	**0.132**	6298 (37.8)	4294 (38.2)	0.007
Inpatient consultation, n (%)	3115 (18.6)	1829 (15.5)	0.081	2906(17.5)	1920 (17.1)	0.010

Abbreviations: ACEi, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; DPP-4i, dipeptidyl peptidase-4 inhibitor; eGFR, estimated glomerular filtration rate; GLP-1RA, glucagon-like peptide-1 receptor agonist; HbA1c, glycated hemoglobin; IPTW, inverse probability of treatment weighting; IQR, interquartile range; MRA, mineralocorticoid receptor antagonist; SD, standard deviation; SGLT-2i, sodium-glucose cotransporter-2 inhibitor; SMD, standardized absolute mean difference; UACR, urinary albumin-to-creatinine ratio.

Note:

[a] eGFR was calculated from creatinine using the CKD-EPI 2021 equation.

[b] UACR was derived directly from lab test results or calculated by dividing urinary albumin by urinary creatinine levels on the same date. Additionally, among patients with missing UACR, we predicted their UACR from their urinary protein-to-creatinine ratio (PCR) or dipstick test results, using validated prediction equations from Sumida et al. (2020).[29]

$$predicted\ UACR\ (from\ PCR) = exp\ (5.2659 + 0.2934 \times log\ (min\ (PCR/50,1)) + 1.5643 \times log\ (max(min(PCR/500,1),0.1)) + 1.1109 \times log\ (max\ (PCR/500,1)) - 0.0773 \times I(female) + 0.0797 \times I(diabetes) + 0.1265 \times I(hypertension))$$

$$predicted\ UACR\ (from\ dipstick) = exp\ (2.0373 + 0.7270 \times I(trace) + 1.6775 \times I(+) + 3.2622 \times I(++) + 4.5435 \times I(> ++) + 0.0822 \times I(female) + 0.27249 \times I(diabetes) + 0.33627 \times I(hypertension)).$$

Table 2. Risk/effect scores, number of events, incidence rate, weighted risk difference and weighted hazard ratio of a composite kidney outcome for GLP-1RA *vs.*DPP-4i, overall and across quarters of risk/effect scores based on risk and effect modelling approaches.

Quarters (median, range) of risk/effect score from,[a] %	**Arm**	**No. of persons**	**No. of events**	**Median follow-up, years [IQR]**	**Incidence rate, events per 1000 PY (95% CI)**	**5-year AR,[b] % (95% CI)**	**5-year RD,[b] % (95% CI)**	**Deviation from average treatment effect, % (95% CI)**	**HR (95% CI)[b]**
Overall	**DPP-4i**	16775	1125	4.2 [2.2-7.1]	13.56 (12.79-14.36)	4.89 (4.47-5.29)	Reference	Reference	Reference
	GLP-1RA	11772	306	2.2 [0.9-4.4]	8.35(7.46-9.31)	4.17 (3.40-4.88)	-0.72 (-1.53 to 0.13)		0.84 (0.72-0.97)
Risk model: KFRE (Performance metrics: C-for benefit = 0.532 [95% CI 0.466-0.599], E_{50}-for-benefit = 0.010 [95% CI 0.004-0.017]**)**									
Q1(lowest), 0.0011 (0.0000-0.0024)	**DPP-4i**	3510	58	5.5 [3.0-8.8]	2.78 (2.15-3.54)	0.85 (0.47-1.30)	Reference		Reference
	GLP-1RA	3627	24	2.3 [0.9-4.8]	2.03 (1.37-2.92)	1.05 (0.38-1.76)	0.20 (-0.58 to 1.04)	0.92 (-0.24 to 1.99)	1.20 (0.63-1.94)
Q2, 0.0044 (0.0024-0.0085)	**DPP-4i**	4025	143	4.9 [2.7-8.3]	6.36 (5.40-7.44)	1.45 (0.94-1.97)	Reference		Reference
	GLP-1RA	3112	45	2.2 [0.9-4.6]	4.51 (3.37-5.92)	1.16 (0.52-1.81)	-0.29 (-1.14 to 0.61)	0.43 (-0.64 to 1.50)	0.79 (0.54-1.16)
Q3, 0.0197 (0.0085-0.0638)	**DPP-4i**	4357	281	4.2 [2.3-7.1]	12.97 (11.54-14.53)	3.80 (3.08-4.63)	Reference		Reference
	GLP-1RA	2779	84	2.3 [1.0-4.2]	9.85 (7.97-	4.19 (2.80-	0.38 (-1.34 to	1.10 (-0.30 to 2.66)	1.06 (0.77-

					12.07)	5.88)	2.26)		1.41)
Q4 (highest), 0.3557 (0.0638-87.4021)	**DPP-4i**	4883	643	3.1 [1.4-5.3]	35.78 (33.12-38.60)	13.80 (12.51-15.03)	Reference		Reference
	GLP-1RA	2254	153	2.1 [0.9-3.9]	24.14 (20.61-28.11)	11.04 (8.57-13.53)	-2.76 (-5.60 to -0.08)	-2.04 (-4.21 to 0.07)	0.70 (0.56-0.86)
Effect model: CSF (Performance metrics: C-for benefit = 0.605 [95% CI 0.464-0.614], E_{50}-for-benefit = 0.009 [95% CI 0.004-0.015])									
Q1 (lowest), -0.1744 (-0.3598 to 0.5889)	**DPP-4i**	4289	111	5.0 [2.7-8.6]	4.53 (3.76-5.41)	0.87 (0.53-1.18)	Reference		Reference
	GLP-1RA	2848	18	1.9 [0.8-3.9]	2.27 (1.44-3.44)	0.64 (0.00-1.47)	-0.23 (-0.97 to 0.65)	0.49 (-0.57 to 1.67)	1.06 (0.54-1.74)
Q2, -0.5156 (-0.6903 to -0.3598)	**DPP-4i**	3860	118	4.8 [2.7-7.7]	5.72 (4.78-6.80)	1.43 (0.98-1.90)	Reference		Reference
	GLP-1RA	3276	25	2.2 [0.9-4.6]	2.43 (1.65-3.47)	1.04 (0.34-1.86)	-0.39 (-1.27 to 0.50)	0.33 (-0.77 to 1.35)	0.90 (0.50-1.35)
Q3, -0.9148 (-1.3333 to -0.6903)	**DPP-4i**	3916	234	4.5 [2.4-7.3]	11.65 (10.25-13.18)	2.93 (2.15-3.72)	Reference		Reference
	GLP-1RA	3221	66	2.5 [1.1-5.0]	6.04 (4.76-7.59)	2.02 (1.09-3.23)	-0.91 (-2.07 to 0.47)	-0.19 (-1.33 to 1.19)	0.70 (0.48-0.96)
Q4 (highest), -2.3729 (-7.0834 to -1.3336)	**DPP-4i**	4710	662	3.1 [1.5-5.3]	37.25 (34.52-40.14)	14.09 (12.85-15.44)	Reference		Reference
	GLP-1RA	2427	197	2.4 [1.1-4.3]	26.11 (22.71-29.88)	12.99 (10.36-15.78)	-1.09 (-3.85 to 1.94)	-0.37 (-2.54 to 2.06)	0.79 (0.65-0.94)

Abbreviations: AR, absolute risk; CI, confidence interval; CSF, causal survival forest; DPP-4i, dipeptidyl peptidase-4 inhibitor; GLP-1RA, glucagon-like peptide-1 receptor agonist; HR, hazard ratio; IQR, interquartile range; KFRE, Kidney Failure Risk Equation; PY, person-years.

Note:

[a] For risk model, risk score refers to the 5-year AR of kidney failure for each patient predicted by KFRE, thus a larger positive value indicates a higher risk of kidney failure; For effect model, effect score refers to the 5-year RD of GLP-1RA *vs.* DPP-4i on composite kidney outcome for each patient predicted by CSF, thus a larger negative value indicates a higher treatment benefit of GLP-1RA on composite kidney outcome.

[b] Weighted ARs, RDs and HRs were estimated using the IPTW method across quarters of risk/effect scores predicted by risk model, and effect model.