Article type: Article

Phenotype-based targeted treatment of SGLT2 inhibitors and GLP-1 receptor agonists in type 2 diabetes

Pedro Cardoso[1], Katie G. Young[1], Anand T.N. Nair[2], Rhian Hopkins[1], Andrew P McGovern[1], Eram Haider[2], Piyumanga Karunaratne[2], Louise Donnelly[2], Bilal A. Mateen[3], Naveed Sattar[4], Rury R. Holman[5], Jack Bowden[1], Andrew T. Hattersley[1], Ewan R. Pearson[2], Angus G. Jones[1], Beverley M. Shields[1], Trevelyan J. McKinley[1†], John M. Dennis[1†*], on behalf of the MASTERMIND consortium

[†]Joint senior

Affiliations:

[1] University of Exeter Medical School. Address: Institute of Biomedical & Clinical Science, RILD Building, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK

[2] University of Dundee. Address: Division of Molecular & Clinical Medicine, Ninewells Hospital and Medical School, University of Dundee, Dundee DD1 9SY, UK

[3] University College London, Institute of Health Informatics. Address: University College London, 222 Euston Rd, London NW1 2DA, London, UK

[4] Address: Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow G12 8TA, UK

[5] Diabetes Trials Unit, Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford. Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford. Address: Diabetes Trials Unit, Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK

Corresponding author

John M Dennis
Institute of Biomedical & Clinical Science, RILD Building, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK
Email: j.dennis@exeter.ac.uk

Tel: +44 (0)7734 940921

Word count: 3057 (excluding abstract and methods)

Keywords: Bayesian non-parametric modelling, heterogeneous treatment effects, precision medicine, type 2 diabetes, SGLT2-inhibitors, GLP1-receptor agonists

Abstract

A precision medicine approach in type 2 diabetes (T2D) could enhance targeting specific glucose-lowering therapies to individual patients most likely to benefit. We utilised Bayesian non-parametric modelling to develop and validate an individualised treatment selection algorithm for two major T2D drug classes, SGLT2-inhibitors (SGLT2i) and GLP1-receptor agonists (GLP1-RA). The algorithm is designed to predict differences in 12-month glycaemic outcome (HbA_{1c}) between the 2 therapies, based on routine clinical features of 46,394 people with T2D in England (27,319 for model development, 19,075 for hold-out validation), with additional external validation in 2,252 people with T2D from Scotland. Routine clinical features, including sex (with females markedly more responsive to GLP1-RA), were associated with differences in glycaemic outcomes. Our algorithm identifies clearly delineable subgroups with reproducible ≥5mmol/mol HbA_{1c} benefits associated with each drug class. Moreover, we demonstrate that targeting the therapies based on predicted glycaemic response is associated with improvements in short-term tolerability and long-term risk of new-onset microvascular complications. These results show that precision medicine approaches to T2D can facilitate effective individualised treatment selection, and that use of routinely collected clinical features could support low-cost deployment in many countries.

Introduction

A precision medicine approach in type 2 diabetes (T2D) would aim to target specific glucose-lowering therapies to individual patients most likely to benefit—for a detailed review of the state of the field, see Dennis, 2020 [1]. Current stratification in T2D treatment guidelines involves preferential prescribing of two major drug classes, SGLT2i-inhibitors (SGLT2i) and GLP1-receptor agonists (GLP1-RA), to subgroups of people with or at high-risk of cardiorenal disease [2]. Evidence informing these recommendations comes from average treatment effect estimates derived from placebo-controlled cardiovascular and renal outcome trials, which have predominantly recruited participants with advanced atherosclerotic cardiovascular risk or established cardiovascular disease [3, 4]. Consequently, there is limited evidence on the benefits of SGLT2i and GLP1-RA for individuals in the broader T2D population and, given the lack of head-to-head trials, of the relative efficacy of the two drug classes for individual patients.

Recent studies have demonstrated a clear potential for a precision medicine approach based on glycaemic response, with the TRIMASTER crossover trial establishing a greater efficacy of SGLT2i compared to another major drug class, DPP4-inhibitors (DPP4i), in those with better renal function, and a greater efficacy of thiazolidinedione therapy compared with DPP4i in those with obesity (BMI >30 kg/m^2) compared to those without obesity [5]. Given these findings, a validated prediction model to support individualised treatment selection has recently been developed for SGLT2i compared with DPP4i therapy using observational and clinical trial data [6]. For GLP1-RA, although recent studies have identified robust heterogeneity in treatment response based on pharmacogenetic markers and markers of insulin secretion [7, 8], the influence of these markers on relative differences in clinical outcomes compared with other drug classes, and therefore their utility for targeting treatment, has not previously been assessed.

Given the lack of evidence to support targeted treatment of SGLT2i compared with GLP1-RA therapies, we aimed to develop and validate a prediction model to provide individualised estimates of differences in 12-month glycaemic (HbA_{1c}) outcomes for the two drug classes. We developed a model using the recently proposed Bayesian Causal Forest (BCF) structure, designed to explicitly identify and predict conditional average treatment effects (CATE), representing differential effects of the two drug classes on HbA1c outcome conditional on the clinical characteristics of individual patients [9]. The BCF framework also minimises confounding from indication bias and allows for flexibility in defining model structure and outputs. Our approach is based on readily-available and routinely collected clinical features, supporting potential low-cost deployment in most countries. We also evaluated the downstream impacts of targeting therapy based on the glycaemic response on

secondary outcomes of weight change, tolerability, and longer-term risk of new-onset microvascular complications, macrovascular complications, and adverse kidney events.

Results

We identified 112,274 study-eligible non-insulin treated people with T2D initiating SGLT2i (n=84,193) or GLP1-RA (n=28,081) for the first time between January 2013 and October 2020 in UK general practice, accessed from Clinical Practice Research Datalink (CPRD) (sFlowchart 1). The mean age of participants was 58.2 (SD=10.9) years, 66,248 (59%) were men, and 88,174 (79%) were of white ethnicity. Baseline clinical characteristics by initiated drug class are listed in Table 1. For the development of the 12-month HbA_{1c} response treatment selection model, individuals with a measured HbA_{1c} outcome were randomly split 60:40 into development (n=31,346) and validation (n=20,865) cohorts (sFlowchart 1; Baseline characteristics by cohort: sTable 1A-B). Mean unadjusted 12-month HbA1c response (change from baseline in HbA1c) was -12.0 (SD 15.3) mmol/mol for SGLT2i and -11.7 (SD 17.6) mmol/mol for GLP1-RA. Specific cohorts were defined for secondary outcomes to maximise the number of included patients for each analysis (sFlowchart 2).

Model development

After variable selection [9] (sFig. 1), the BCF model identified multiple clinical factors predictive of HbA_{1c} response with SGLT2i (the reference drug class in the model) henceforth referred to as the prognostic factors, and multiple factors predictive of *differential* HbA_{1c} response with GLP1-RA compared to SGLT2i therapy (henceforth referred to as the differential factors —Table 2). The final BCF model was fitted to 27,319 (87.2%) individuals with complete data for all selected clinical factors. Overall model fit and performance statistics for predicting achieved HbA_{1c} outcome in internal validation for both the development and hold-out cohorts are reported in sTable 2. Model predictions were similar whether or not we incorporated a propensity score into the BCF model (as recommended by Hahn *et al.* [10]) (sFig. 2). Therefore, we removed the propensity score covariate from the final model to support easier future implementation of the model within clinical practice.

In the development cohort, the predicted CATE was a 0.1 (95%CI -0.3;0.5) mmol/mol benefit with GLP1-RA over SGLT2i, suggesting similar average efficacy of both therapies. However, across individuals, there was marked heterogeneity in predicted CATE (Fig. 1A), with the model predicting an average HbA_{1c} benefit on SGLT2i therapy for 13,110 (48%) individuals and on GLP1-RA for 14,209 (52%) individuals. 4,787 (17.5%) of the development cohort had a predicted HbA1c benefit >3 mmol/mol (3mmol/mol is used widely as minimally important difference in clinical trials) with SGLT2i over GLP1-RA, and 5,551 (20.3%) had a predicted HbA_{1c} benefit >3 mmol/mol with GLP1-RA over SGLT2i.

Model calibration

Calibration by decile of model-predicted CATE estimates was good in the development cohort (n=27,319; Fig. 1B), the hold-back CPRD validation cohort (n=19,075, Fig. 1C), and in propensity-matched cohorts (sFig. 3). In practice, true CATE estimates are unobserved since a single individual receives only one therapy, meaning the counterfactual outcome they would have had on the alternative therapy is unobserved. Therefore, for each decile calibration was based on comparing mean predicted CATE estimates to the mean differences in outcome HbA1c of individuals receiving SGLT2i compared to GLP1-RA—for full details, see Methods.

We also evaluated model performance in an external Scottish cohort—Tayside & Fife (n=2,252 [1,837 initiating SGLT2i, 415 initiating GLP1-RA]; Baseline characteristics: sTable 1C). A similar distribution of predicted CATE to CPRD was observed (Extended Fig. 1A). Due to the smaller cohort size, calibration was assessed by quintile of predicted CATE, with a clear difference between upper (favouring GLP1-RA) and lower (favouring SGLT2i) quintiles, but modest calibration in middle quintiles (Extended Fig. 1B). Among 81 (3.6%) individuals with a model-predicted HbA_{1c} benefit >5 mmol/mol for SGLT2i over GLP1-RA, there was a 7.4 mmol/mol (95%CI 0.1;14.8) benefit for SGLT2i (Extended Fig. 1C). In contrast, among 150 (6.7%) individuals with a model-predicted HbA_{1c} benefit >5 mmol/mol for GLP1-RA over SGLT2i, there was a 5.6 mmol/mol (95%CI -0.9;12.1) benefit on GLP1-RA.

Model interpretability

Stratifying the combined development and validation cohorts with complete predictor data (n=46,394) into subgroups defined by predicted CATE, there were clear differences in clinical characteristics, with those having a greater predicted HbA1c benefit with GLP1-RA over SGLT2i being predominantly female and older, with lower baseline HbA_{1c}, eGFR and BMI (Fig. 2, sTable 1E). SGLT2i were predicted to have a greater HbA_{1c} benefit over GLP1-RA for 32% of those with baseline HbA_{1c} levels <64, 39% 64-75, 54% 75-86, and 67% ≥86 mmol/mol. Given their average older age, those with the greatest (>5 mmol/mol) predicted benefit on GLP1-RA also had a higher prevalence of microvascular complications (75.2% versus 40.8%), cardiovascular disease (36.5% versus 15.8%), chronic kidney disease (25.1% versus 0.9%), and heart failure (8.8% versus 4.2%) (sTable 1E). An evaluation of relative variable importance identified the number of other current glucose-lowering drugs (a higher number of concurrent therapies favouring SGLT2i as the optimal treatment), sex, current age, and to a lesser extent BMI and HbA_{1c} as the most influential predictors (relative importance ≥3%). In contrast, microvascular complications and cardiovascular comorbidities had very modest effects on differential response (Extended Fig. 2).

Replication of sex differences in glycaemic response in clinical trials

Whilst previous analyses of clinical trials and observational data for SGLT2i have shown a greater HbA_{1c} response in males compared to females, which we additionally reproduced in Tayside & Fife (Extended Fig. 3), sex differences in GLP1-RA response have not been clearly established. Here, we focused on individual-level randomised clinical trial data of GLP1-RA from the HARMONY programme (Liraglutide [n=389] and Albiglutide [n=1,682]), [11], the PRIBA prospective cohort study (Liraglutide [n=397], exenatide [n=223]) [7], and Tayside & Fife (n=415). Baseline characteristics for the cohorts are reported in sTable 1C-D. Across all studies, there was consistent evidence of a greater baseline HbA_{1c} adjusted glycaemic response in females versus males; this was most marked for Liraglutide in the HARMONY 7 trial [11] where a 4.4 (95%CI 2.2;6.3) mmol/mol greater response in females than males was observed.

Effect of targeting therapy based on predicted HbA_{1c} outcome on other short- and long-term outcomes

To assess the potential associations of targeting treatment based on predicted HbA_{1c} benefit with other short- and long-term outcomes, we divided the study population into six subgroups based on clinically relevant differences in predicted CATE (HbA1c differences of 0-3, 3-5 and >5 mmol/mol between therapies). We then compared average differences in each short and long-term outcome in individuals receiving SGLT2i compared to GLP1-RA within each subgroup, for adjusted (Fig. 3), propensity-matched (sFig. 4), and double robust adjusted models (sFig. 5). Specific subpopulations were defined for each short-term outcome based on the availability of observed outcome data (12-month HbA_{1c} change from baseline [to evaluate absolute response] n=87,835; 12-month weight change n=41,728; treatment discontinuation within 6 months [a proxy for tolerability] n=77,741) (sFlowchart 2). Longer-term outcomes were evaluated up to 5 years from drug initiation, excluding individuals with a history of cardiovascular disease or chronic kidney disease for major adverse cardiovascular event (MACE), heart failure, and adverse kidney (composite of ≥40% decline in eGFR or kidney failure [12]) outcomes (n=52,052) and individuals with a history of retinopathy, neuropathy and nephropathy for microvascular outcome (n=34,524). (sFlowchart 2).

For HbA_{1c} change from baseline, of the 6,856 individuals (7.8%) with a predicted HbA_{1c} benefit on SGLT2i of ≥5 mmol/mol, those who received SGLT2i had on average a 23.3 mmol/mol (95%CI 22.6;24.0) reduction in HbA_{1c} and those who received GLP1-RA had on average an 18.4 mmol/mol (95%CI 17.6;19.3) reduction in HbA_{1c} (Fig. 3A.1). In contrast, 7,293 individuals (8.3%) with a predicted HbA_{1c} benefit on GLP1-RA of ≥5 mmol/mol, those receiving GLP1-RA had on average a 15.7 mmol/mol (95%CI 14.8;16.6) reduction in HbA_{1c}, and those receiving SGLT2i had on average a 9.0 mmol/mol (95%CI 8.2;9.7) reduction in HbA_{1c}. Similar differences between subgroups were seen

when stratified by major UK self-reported ethnicity groups (White and Non-White [South Asian, Black, Other or Mixed]) (Extended Fig. 4), and by history of cardiovascular disease (sFig. 10).

Observed weight change was consistently greater for individuals treated with SGLT2i compared to GLP1-RA across all subgroups (Fig. 3A.2). Short-term discontinuation was lower in those treated with the drugs predicted to have the greatest glycaemic benefit by the model, mainly reflecting differences in SGLT2 discontinuation across predicted levels of differential glycaemic response (Fig. 3A.3). Relative risk of new-onset microvascular complications also varied by subgroup, with a lower risk with SGLT2i versus GLP1-RA only in subgroups predicted to have a glycaemic benefit with SGLT2i (Fig. 3B.1). HRs for the risk of new-onset MACE were similar overall (HR 1.02 [95%CI 0.89;1.18]) and by subgroup (Fig. 3B.2). HRs for the risks of both new-onset heart failure and adverse kidney outcomes were lower with SGLT2i (heart failure HR 0.71 [95%CI 0.59;0.85]; CKD HR 0.41 [95%CI 0.30;0.56]) with no clear evidence of a difference by subgroup (Fig. 3B.3, sFig. 6C).

Comparison of model predictions with our previously published treatment selection model for SGLT2i and DPP4i therapies

In n=82,933 eligible patients, predictions for HbA_{1c} response with SGLT2i from the SGLT2i v GLP1-RA treatment selection model were highly concordant ($R^2 > 0.92$) with those from our recently published SGLT2i versus DPP4i treatment selection model, which was developed in an independent UK primary care cohort [6] (sFig. 7). Estimating differential HbA_{1c} responses using both models in our study population with complete data suggested SGLT2i is the predicted optimal therapy for HbA_{1c} in 48.2% (n=39,975) of individuals, GLP1-RA the predicted optimal therapy in 51.3% (n=42,519), and DPP4i the optimal therapy for only 0.5% (n=439).

Prototype treatment selection model

A prototype treatment selection model app (provided for research purposes only) providing individualised predictions of differences (with uncertainty) in HbA1c outcomes is available at: https://pml204.shinyapps.io/SGLT2_GLP1_calculator/

Discussion

We have developed a novel treatment selection algorithm using state-of-the-art Bayesian methods to predict differences in one-year glycaemic outcomes for SGLT2i and GLP1-RA and validated the algorithm in a hold-out cohort and an independent external validation cohort. Our individualised precision medicine approach shows glycaemic response-based targeting of these two major drug classes in people with T2D can not only optimise glycaemic control, but is also associated with improved tolerability and reduced risk of new-onset microvascular complications. In contrast, we found limited evidence for heterogeneity in other clinical outcomes, with overall equipoise between the two therapies for new-onset MACE and a clear overall benefit with SGLT2i over GLP1-RA for new-onset heart failure and adverse kidney outcomes (≥40% eGFR decline or renal failure). Predictions are based on individual patient-level routine clinical characteristics, meaning the model can be deployed in most countries worldwide without the need for additional testing.

Our approach differs from notable recent studies that have attempted to sub-classify people with T2D or used dimensionality reduction to represent T2D heterogeneity [13, 14, 15]. Whilst these alternative approaches can provide important insight into underlying heterogeneities of T2D patient characteristics, they, by definition, lose information about the specific characteristics of individual patients, meaning they are sub-optimal for accurately predicting the treatment or disease progression outcomes for individuals [16]. If subclassification approaches based on clinical features are to have potential clinical utility, they will need to be updated over time as an individual's phenotype evolves [17]. In contrast, our 'outcomes-based' approach enables the prediction of optimal therapy when a treatment decision is made, uses the specific information available for a patient at that point in time and avoids sub-classification.

Although BCF models are only causal under specific assumptions [18], they can provide insights into differences in the underlying mechanisms of action of GLP1-RA and SGLT2i, and the clinical utility of these differences. The strongest predictor of a differential glycaemic response was the number of currently prescribed glucose-lowering therapies, which is a likely proxy of the degree of diabetes progression (and, therefore, underlying beta cell failure) of an individual. A plausible biological explanation for the relevance of this proxy is that an attenuated GLP1-RA response has been observed in individuals with more severe diabetes; previous clinical studies have identified that markers of beta cell failure (including longer diabetes duration, insulin co-treatment and lower fasting c-peptide) are associated with a lower glycaemic response to GLP1-RA [7], with no evidence of differences for SGLT2i [19]. The favouring of GLP1-RA over SGLT2i in females is novel but is supported by our trial validation and recent pharmacokinetic data demonstrating higher circulating GLP1-RA drug concentrations and, consequently, greater HbA1c reduction in females compared with

males [20]. For SGLT2i, increased urinary glucose excretion likely explains the greater relative glycaemic efficacy with higher baseline HbA_{1c} and eGFR, which in concordance with our analysis, has been previously demonstrated in trial data [6]. Of note, the comorbidities included in the final model had modest effects on HbA1c and are likely to be proxy measures of factors underlying differential response to these therapies.

Our study represents the second application of our novel validation framework for precision medicine models, which, in the absence of true observed outcomes (for an individual patient on one therapy, the counterfactual outcome they would have had on an alternative therapy cannot be observed [21]), evaluates accuracy in subgroups defined by predicted CATE. The previous study developed a treatment selection model for SGLTi2 versus DPP4i therapy in a separate dataset, with models developed in primary care data validating well in head-to-head clinical trial datasets. Although this demonstrated marked heterogeneity in the relative glycaemic outcome, most (84%) individuals had a greater glycaemic reduction with SGLT2i. In contrast, this GLP1-RA/SGLT2i model shows even greater heterogeneity in treatment effects but with equipoise on average treatment effects between the two therapies (52% favouring GLP1-RA). Furthermore, we demonstrate that optimising therapy based on predicted glycaemic response may lower microvascular complication risk, a finding concordant with evidence from the UKPDS study on the importance of good glycaemic control to lower the risk of microvascular disease [22, 23].

Further developments to this model could include the incorporation of non-routine and pharmacogenetic markers (recently identified for GLP1-RA) [8], and additional glucose-lowering drug classes, in particular, off-patent sulfonylureas and pioglitazone to support the deployment of the algorithm in lower-income countries where the availability of newer medications may be limited. Assessment of semaglutide, a GLP-1RA with potent glycaemic effect excluded here due to low numbers prescribed during the period of data availability, and tirzepatide, a dual GIP and GLP-1 receptor agonist not currently available in the UK, is an important area for future research as our model may benefit from recalibration for these newer therapies. Although our ethnicity-specific validation suggests good performance in people of non-white ethnicity, setting and ethnicity-specific validation and optimisation would also improve future clinical utility. Given the possibility of selection bias due to non-random treatment assignment, validation in a dataset where individuals were randomised to therapy would further strengthen the evidence for model deployment. However, few active comparator trials of these two therapies have been conducted [24], and to our knowledge, none are available for data-sharing. Ultimately, research, likely in even larger datasets, is needed on whether individualised models for other short-and-long term outcomes beyond glycaemia, particularly cardiorenal disease, can further improve current prescribing approaches [25].

Finally, a limitation of our study is that despite being state-of-the-art and with a key advantage of allowing estimation of predictions with uncertainty, and so facilitating more transparent evaluation, the Bayesian causal forest methods we applied are subject to ongoing development in several key areas such as variable selection [9, 26], scalability, and handling of missing data [27].

In conclusion, our study demonstrates a clear potential for targeted prescribing of GLP1-RA and SGLT2i to individual people with T2D based on their clinical characteristics to improve glycaemic outcomes, tolerability, and risk of microvascular complications. This provides an important advance on current T2D guidelines, which only recommend preferentially prescribing these therapies to individuals with, or at high risk of, cardiorenal disease, with no clear evidence to choose between the two drug classes. Precision T2D prescribing based on routinely-available characteristics has the potential to lead to more informed and evidence-based decisions on treatment for people with T2D worldwide in the near future.

Methods

Study population

Patients with type 2 diabetes initiating SGLT2i and GLP1-RA therapies between 1st January 2013 and 31st October 2020 were identified in the UK Clinical Practice Research Datalink (CPRD) Aurum dataset [28], following our previously published cohort profile [29] (see GitHub: Exeter-Diabetes/CPRD-Codelists for the codelists used). We excluded patients that were prescribed either therapy as first-line treatment (not recommended in UK guidelines) [30], those co-treated with insulin, and those with a diagnosis of end-stage renal disease (ERSD) (sFlowchart 1). Due to the low numbers of eligible patients we also excluded individuals initiating the GLP1-RA semaglutide (n=784 study eligible with outcome HbA1c recorded)) [31]. This final CPRD cohort was randomly split 60:40 into development and hold-back validation sets, maintaining the proportion of individuals receiving SGLT2i and GLP1-RA in each set.

Additional cohorts

The same eligibility criteria were applied to define an independent population in Scotland for model validation (SCI-Diabetes [Tayside & Fife] [32], containing longitudinal observational data on biochemical investigations and prescriptions). To assess reproducibility of differences in HbA_{1c} response by sex with GLP1-RA therapy, we accessed individual-level data on participants initiating the GLP1-RAs Albiglutide and Liraglutide in the HARMONY clinical trial programme, an international randomised placebo-controlled trial designed to evaluate the cardiovascular benefit of Albiglutide with type 2 diabetes [11], and the PRIBA prospective cohort study (United Kingdom 2011-2013) [7], designed to test whether individuals with low insulin secretion have lesser glycaemic response to incretin-based treatments.

Outcomes

The primary outcome was achieved HbA_{1c} at twelve months post-drug initiation for individuals who remained on unchanged glucose-lowering therapy. Given the variability in the timing of follow-up testing in UK primary care, this outcome was defined as the closest eligible HbA_{1c} value to 12 months (within 3–15 months) after initiation to maximise sample size. To allow for potential differential effects of follow-up duration on HbA_{1c}, we included an additional covariate to capture the month post-initiation that the HbA_{1c} value was recorded.

Secondary outcomes comprised short-term outcomes: change in weight 12 months after initiation (closest recorded weight to 12 months, within 3–15 months), and, as a proxy for drug tolerability, treatment discontinuation within 6 months of drug initiation (as such short-term discontinuation is unlikely to be related to a lack of glycaemic response); and longer-term outcomes up to 5-years after

initiation: new-onset major adverse cardiovascular events (MACE: composite of myocardial infarction, stroke and cardiovascular death); new-onset heart failure; and new-onset adverse kidney outcome (a drop of ≥40% in eGFR from baseline or reaching CKD stage 5 [12]); and new-onset microvascular complications (see sFlowchart 2).

Predictors

Candidate predictors comprised — current age, duration of diabetes, year of therapy start, sex, ethnicity (major UK groups: White, South Asian, Black, mixed, other ethnicities), social deprivation (index of multiple deprivation quintile), smoking status, the number of current, and ever, prescribed glucose lowering drug classes, baseline HbA_{1c} (closest to treatment start date; range in previous 6 months to +7 days), and other measured clinical features: BMI, eGFR (using the Chronic Kidney Disease Epidemiology Collaboration formula [33]), HDL cholesterol, alanine aminotransferase, albumin, bilirubin, total cholesterol and the mean arterial blood pressure (defined as closest values to treatment start in the previous two years), microvascular complications: nephropathy, neuropathy, retinopathy , and major comorbidities: angina, atherosclerotic cardiovascular disease, atrial fibrillation, cardiac revascularisation, heart failure, hypertension, ischaemic heart disease, myocardial infarction, peripheral arterial disease, stroke, transient ischaemic attack, chronic kidney disease and chronic liver disease.

Treatment selection model development

Model overview

We aimed to build a flexible Bayesian treatment selection model [10] to predict differential glycaemic response that can be readily deployed in clinical practice. The model development process consisted of a first step of propensity score estimation and a second step of developing a treatment selection model using a Bayesian additive regression tree (BART) framework [9, 26]. We aimed to balance predictive accuracy with model parsimony and build the model around a limited number of routinely collected variables, thus facilitating its use in clinical practice. Individuals were excluded from the development and validation sets if they initiated multiple glucose-lowering treatments on the same day; their therapies were initiated less than 61 days since the start of a previous therapy; their baseline HbA_{1c} was <53 mmol/mol; they had a missing baseline HbA_{1c}; or they had a missing outcome HbA_{1c} (sFlowchart 1).

The treatment selection model was developed using Bayesian Causal Forests (BCF) [10], a framework specifically designed to estimate heterogeneous treatment effects. The BCF approach places flexible BART priors on the mean functions relating to a prognostic component of the model (representing outcomes in the reference group, in this case, SGLT2i) but also a moderator component that

operates over-and-above the prognostic component (in this case, the differential treatment effect of GLP1-RA relative to SGLT2i) [10, 34]. With this method, the shrinkage applied to the differential treatment effects can be adjusted independently of the prognostic effects. We used the *bcf* [10] and *sparseBCF* [26] packages in R to fit the model(s) using Markov chain Monte Carlo (MCMC). By default, *bcf* places stronger regularisation on the moderator side of the model, shrinking the treatment effects towards homogeneity where there is a lack of strong evidence to the contrary. For this model, we are comparing two therapies against each other (rather than a therapy to a control group), so we used the same prior structure for the prognostic and moderator parts of the model. A full description of the model parameters and priors is included in the Supplementary Materials. Currently, the standard BCF software cannot account for missing data [35], so we used a complete case analysis, informed by our previous study showing a limited impact of missing data on predicting CATE in a similar primary care dataset [27].

Propensity score estimation

The BCF developers [10] recommend including a propensity score variable in the prognostic component of BCF models to help alleviate regularisation-induced confounding due to prescribing by indication [10]. We used a standard BART model [35] for the propensity score, fitted using MCMC and the *bartMachine* package in R [35]. All variables extracted initially in the development cohort were used for initial model fitting. However, a subset of the most predictive variables was selected by applying a threshold defined by the proportion of times each predictor was chosen as a spitting rule divided by the total number of splitting rules appearing in the model [35, 36] (sFig. 8). The propensity score model was then refitted with the selected variables. To assess convergence, we monitored the available parameters according to the guidance provided by Kapelner *et al*. [35] and Gelman-Rubin $\hat{R}$ values. A description of the model priors is included in the Supplementary Material. The BART propensity score model converged quickly, so we ran 25,000 iterations with the first 15,000 discarded for burn-in; trace plots are available on request and $\hat{R} < 1.02$. To assess the performance of the final model, received operating characteristic (ROC) and precision-recall curves were fitted to both the development and validation cohorts (sFig. 9).

Variable selection

Variable selection was deployed to develop a parsimonious final model whilst maintaining predictive accuracy. To do this, we used a two-stage approach, wherein the first stage, we built a sparse BCF model [9] incorporating all candidate predictors. Sparse BCF extends standard BCF by replacing the uniform prior distribution placed over the splitting probabilities of each variable (which means that by default, each variable has the same prior probability of being selected for splitting) with a uniform Dirichlet prior over the splitting probabilities. As the model converges, the posterior distribution for

these splitting probabilities induces sparsity by assigning higher weight and, thus, higher variable importance to more predictive covariates. This prior was used in the model's prognostic and moderator parts [9]. To define the final predictor set, we selected only variables with a posterior mean splitting probability greater than 1/number of variables. This was a subjective choice, but one we found was sufficient to minimise the number of final predictors without meaningfully affecting predictive accuracy. We ran the sparse BCF model for 250,000 iterations, discarding the first 200,000 iterations as burn-in and monitoring convergence using trace plots (available on request) and Gelman-Rubin $\hat{R}$ values (where all $\hat{R} < 1.01$). A description of the model parameters and priors is included in the Supplementary Material.

Final model fit

Following variable selection, we fitted a final model using standard BCF without sparsity-inducing priors (since individual-level predictions are not currently possible from the sparse BCF software). The model used 300,000 iterations, discarding the first 200,000 iterations and thinning the remaining iterations by 4, resulting in 25,000 final posterior samples. The propensity score was not included in the final predictor set as it did not meet our threshold for variable selection. As a sensitivity analysis, we refitted the model including the propensity score in the predictor set and compared predictions across the two models (sFig. 11).

Variable importance (based on the best linear projection)

Given the known challenge of extracting variable importance from tree-based models, we implemented a pseudo-variable importance measure defined as the proportion of R^2 associated with each variable for predicting the CATE [37]. This was estimated from a linear regression model using all selected variables for the differential part of the model as predictors (with continuous predictors fitted as 3-knot restricted cubic splines) and the predicted CATE as the outcome [38]. To assess how CATE estimates varied across major routine clinical features, we also summarised the marginal distributions of sex, baseline HbA_{1c}, eGFR, current age, and BMI across subgroups defined by the degree of predicted glycaemic differences (SGLT2i benefit of 0–3, 3–5 or ≥5 mmol/mol; GLP1-RA benefit of 0–3, 3–5 or ≥5mmol/mol).

Model validation

Evaluating the accuracy of predicted CATE is a significant challenge since each individual only has outcome data for the treatment they initiated but not for the competing treatment (which they did not take) [21], meaning the treatment effect is not observed at the individual level (it is a so-called *counterfactual* effect). As such, to validate predicted CATE estimates, we first split validation sets into sub-groups based on their predicted CATE estimates from the BCF model and then compared the *average* CATE estimate within each sub-group to estimates derived from a set of alternative

models fitted to each of the sub-groups in turn. These latter models target the *average treatment effect* (ATE) within a population of individuals (rather than the conditional average treatment effect [CATE]), with desirable properties justified in the literature [6]. An earlier version of this validation framework is described previously [6]. If the average CATE estimates in each sub-group (from the BCF model) align with the ATE estimates from the alternative models, then this provides evidence that the average treatment effect estimates are consistent across different inference methods within each sub-group. Restricting the ATE estimates for each sub-group allows for simpler comparison ATE models to be used, since the distribution of covariates in each sub-group is expected to be more consistent within each subgroup than for the complete data. For validation, subgroups were defined by decile of predicted CATE in CPRD and, due to lower patient numbers, by quintile in the Tayside & Fife cohort. We then employed three different approaches to estimate the ATE within each subgroup:

Regression adjustment (primary approach): We included all patients and used Bayesian linear regression to estimate the ATE within each subgroup, adjusting for the full covariate set used in the HbA1c treatment selection model (Table 2), with all continuous predictors included as 3-knot restricted cubic splines [39].

Propensity score matching: Individuals receiving each drug class within each subgroup were matched by propensity score (the same propensity score used during HbA1c model development), using a caliper distance of 0.05, no replacement and in decreasing order of propensity score values. After defining this restricted patient subset, unadjusted linear regression models were used to estimate the ATE within each subgroup [40].

Propensity score matching with adjustment: The linear regression models of approach 2 were refitted using a double robust approach by adjusting for the full covariate set used in the HbA1c treatment selection model (Table 2).

In sensitivity analysis, we further assessed the accuracy of predicted HbA_{1c} treatment effects in those of non-white ethnicity and in those with and without cardiovascular disease. We also evaluated the reproducibility of observed differences in HbA1c response by sex in participants receiving GLP1-RA in the HARMONY clinical trial, the PRIBA prospective study, and Tayside & Fife.

Secondary outcomes

Specific cohorts were defined to evaluate each secondary outcome to mitigate selection bias and maximise the number of individuals available for analysis (sFlowchart 2). All cohorts required complete predictor data for the HbA1c-based treatment selection model. To evaluate treatment effect heterogeneities, subgroups were defined by the degree of predicted glycaemic differences

(SGLT2i benefit of 0–3, 3–5 or ≥5 mmol/mol; GLP1-RA benefit of 0–3, 3–5 or ≥5mmol/mol). As for validation of differences in HbA_{1c} outcomes, three approaches were employed to estimate CATE: regression adjustment (primary approach), propensity score matching, and propensity score matching with predictor variable set adjustment (propensity score model was refitted, following the procedure mentioned before, with the additional inclusion of baseline cardiovascular risk as a predictor (QRISK2 predicted probability of new-onset myocardial infarction or stroke [41]).

For 12-month weight change, we included all patients with a recorded baseline weight (closest value to 2 years prior to treatment initiation) and a valid outcome weight. Treatment effects were estimated using a Bayesian linear regression model with an interaction between the received treatment and the predicted HbA_{1c} treatment benefit subgroup, with adjustment for baseline weight. We included all patients initiating therapy for treatment discontinuation and estimated CATE using Bayesian logistic regression with a treatment-by-HbA_{1c} benefit subgroup interaction. For longer-term outcomes, we included only individuals without the outcome of interest at therapy initiation, thus evaluating only incident events. Patients were followed for up to 5 years using an intention-to-treat approach from the date of therapy initiation until the earliest of: the outcome of interest, the date of GP practice deregistration or death, or the end of the study period. For each outcome, Bayesian Cox proportional hazards models with treatment-by-HbA1c benefit subgroup interactions were fitted with additional adjustment for QRISK2 predicted probability of new-onset myocardial infarction or stroke. A description of priors for each model is included in the Supplementary Material. The models used 10,000 MCMC iterations, discarding the first 5,000 iterations as burn-in for four chains. Convergence was monitored using trace plots and Gelman-Rubin $\hat{R}$ values. Here all $\hat{R} < 1.04$ and trace plots are available on request.

All analyses were conducted using R (version 4.1.2). We followed TRIPOD prediction model reporting guidance (see Supplementary Materials—TRIPOD checklist) [42].

Acknowledgements: This article is based in part on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare products Regulatory Agency. CPRD data is provided by patients and collected by the NHS as part of their care and support. Approval for CPRD data access and the study protocol was granted by the CPRD Independent Scientific Advisory Committee (eRAP protocol number: 22_002000). This publication is based in part on research using data from GSK that has been made available through secured access. GSK has not contributed to or approved, and is not in any way responsible for, the contents of this publication. The PRIBA study was funded by A National Institute for Health Research (U.K.) Doctoral Research Fellowship (DRF-2010-03-72, AGJ) and supported by the National Institute for Health Research Clinical Research Network. The authors thank the members of the Predicting Response to Incretin Based Agents (PRIBA) study group and all cohort participants. ATH and BMS are supported by the NIHR Exeter Clinical Research Facility; the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. PC, KGY, JB, TJM and JD are supported by Research England's Expanding Excellence in England (E3) fund.

Author contributions: PC, JMD, BS, TJM, ATH, AGJ and ERP conceived and designed the study. PC, with support from JMD, BS and TJM analysed the data and developed the code. KGY, RH and AM helped with curating the CPRD dataset. ATNN, PK, EH and LD helped analyse the Scottish independent dataset. All authors contributed to the writing of the article, provided support for the analysis and interpretation of results, critically revised the article, and approved the final article.

Funding: This research was funded by the Medical Research Council (UK) (MR/N00633X/1) and a BHF-Turing Cardiovascular Data Science Award (SP/19/6/34809).

Conflict of interest: APM declares previous research funding from Eli Lilly and Company, Pfizer, and AstraZeneca. BAM holds an honorary post at University College London for the purposes of carrying out independent research, and declares payments to their institution from the MRC, HDRUK and BHF. NS declares personal fees from Abbott Diagnostics, Afimmune, Amgen, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Hanmi Pharmaceuticals, Merck Sharp & Dohme, Novartis, Novo Nordisk, Pfizer and Sanofi and grants to his University from AstraZeneca, Boehringer Ingelheim, Novartis and Roche Diagnostics. RRH reports research support from AstraZeneca, Bayer and Merck Sharp & Dohme, and personal fees from Anji Pharmaceuticals, Bayer, Novartis and Novo Nordisk. JB is an employee of Novo Nordisk, outside of the submitted work. ERP has received honoraria for speaking from Lilly, Novo Nordisk and Illumina. AGJ has received research funding from the Novo Nordisk foundation. Representatives from GSK, Takeda, Janssen, Quintiles, AstraZeneca and Sanofi attend meetings as part of the industry group involved with the MASTERMIND consortium. No industry representatives were involved in the writing of the manuscript or analysis of data. For all authors these are outside the submitted work; there are no other relationships or activities that could appear to have influenced the submitted work.

Code availability statement: All R code used for the analysis is provided at https://github.com/Exeter-Diabetes/CPRD-Pedro-SGLT2vsGLP1.

Data availability statement: The UK routine clinical data analysed during the current study are available in the CPRD repository (CPRD; https://cprd.com/research-applications), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. For re-using these data, an application must be made directly to CPRD. Data from Scotland are anonymized real-world medical records available by request through the Scottish Care Information-Diabetes Collaboration, Tayside & Fife, Scotland unit (https://www.sci-diabetes.scot.nhs.uk/). Clinical trial data are not publicly available, for access an application must be made directly to GSK and www.ClinicalStudyDataRequest.com

References

[1] J. M. Dennis, "Precision medicine in type 2 diabetes: using individualized prediction models to optimize selection of treatment," *Diabetes,* vol. 69, no. 10, pp. 2075-2085, 2020.

[2] M. J. Davies, V. R. Aroda, B. S. Collins, R. A. Gabbay, J. Green, N. M. Maruthur, S. E. Rosas, S. D. Prato, C. Mathieu, G. Mingrone, P. Rossing, T. Tankova, A. Tsapas and J. B. Buse, "Management of hyperglycemia in type 2 diabetes, 2022. A consensus report by the american diabetes association (ADA) and the european association for the study of diabetes (EASD)," *Diabetes Care,* vol. 45, no. 11, p. 2753–2786, 2022.

[3] W. T. Cefalu, S. Kaul, H. C. Gerstein, R. R. Holman, B. Zinman, J. S. Skyler, J. B. Green, J. B. Buse, S. E. Inzucchi, L. A. Leiter, I. Raz, J. Rosenstock and M. C. Riddle, "Cardiovascular outcomes trials in type 2 diabetes: where do we go from here? Reflections from a diabetes care editors' expert forum," *Diabetes Care,* vol. 41, no. 1, pp. 14-31, 2018.

[4] A. McGovern, M. Feher, N. Munro and S. de Lusignan, "Sodium-glucose co-transporter 2 (SGLT2) inhibitor: comparing trial data and real-world use," *Diabetes Therapy,* vol. 8, pp. 365-376, 2017.

[5] B. M. Shields, J. M. Dennis, C. D. Angwin, F. Warren, W. E. Henley, A. J. Farmer, N. Sattar, R. R. Holman, A. G. Jones, E. R. Pearson, A. T. Hattersley and TriMaster Study group, "Patient stratification for determining optimal second-line and third-line therapy for type 2 diabetes: the TriMaster study," *Nature Medicine,* vol. 29, pp. 376-383, 2023.

[6] J. M. Dennis, K. G. Young, A. P. McGovern, B. A. Mateen, S. J. Vollmer, M. D. Simpson, W. E. Henley, R. R. Holman, N. Sattar, E. R. Pearson, A. T. Hattersley, A. G. Jones and B. M. Shields, "Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in people with type 2 diabetes: a retrospective cohort study," *The Lancet Digital Health,* vol. 4, no. 12, pp. 873-883, 2022.

[7] A. G. Jones, T. J. McDonald, B. M. Shields, A. V. Hill, C. J. Hyde, B. A. Knight, A. J. Hattersley and for the PRIBA Study group, "Markers of β-cell failure predict poor glycemic response to GLP-1 receptor agonist therapy in type 2 diabetes," *Diabetes Care,* vol. 39, no. 2, pp. 250-257, 2016.

[8] A. Y. Dawed, A. Mari, A. Brown, T. J. McDonald, L. Li, S. Wang, M.-G. Hong, S. Sharma, N. R. Robertson, A. Mahajan, X. Wang, M. Walker, S. Gough, L. M. Hart, K. Zhou, I. Forgie, H. Ruetten, I. Pavo, P. Bhatnagar, A. G. Jones, E. R. Pearson and for the DIRECT consortium, "Pharmacogenomics of GLP-1 receptor agonists: a genome-wide analysis of observational data and large randomised controlled trials," *The Lancet Diabetes & Endocrinology,* vol. 11, no. 1, pp. 33-41, 2023.

[9] A. Caron, G. Baio and I. Manopoulou, "Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation," *Journal of Computational and Graphical Statistics,* vol. 31, no. 4, pp. 1202-1214, 2021.

[10] P. R. Hahn, J. S. Murray and C. M. Carvalho, "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion)," *Bayesian Analysis,* vol. 15, no. 3, pp. 965-1056, 2020.

[11] R. E. Pratley, M. A. Nauck, A. H. Barnett, M. N. Feinglos, F. Ovalle, I. Harman-Boehm, J. Ye, R. Scott, S. Johnson, M. Stewart, J. Rosenstock and HARMONY 7 study group, "Once-weekly albiglutide versus once-daily liraglutide in patients with type 2 diabetes inadequately controlled on oral drugs (HARMONY 7): a randomised, open-label, multicentre, non-inferiority phase 3 study," *Lancet Diabetes Endocrinology,* vol. 2, no. 4, pp. 289-297, 2014.

[12] M. E. Grams, N. J. Brunskill, S. H. Ballew, Y. Sang, J. Coresh, K. Matsushita, A. Surapaneni, S. Bell, J. J. Carrero, G. Chodick, M. Evans, H. J. Heerspink, L. A. Inker, K. Iseki, P. A. Kalra, H. L. Kirchner, B. J. Lee, A. Levin, R. W. Major, J. Medcalf, G. N. Nadkarni, D. M. Naimark, A. C. Ricardo, S. Sawhney, M. M. Sood, N. Staplin, N. Stempniewicz, B. Stengel, K. Sumida, J. P. Traynor, J. v. d. Brand, C.-P. Wen, M. Woodward, J. W. Yang, A. Y.-M. Wang, N. Tangri and C. P. C. , "Development and validation of prediction models of adverse kidney outcomes in the population with and without diabetes," *Diabetes Care,* vol. 45, no. 9, pp. 2055-2063, 2022.

[13] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spégel, H. Mulder and L. Groop, "Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables," *The Lancet Diabetes & Endocrinology,* vol. 6, no. 5, pp. 361-369, 2018.

[14] A. T. N. Nair, A. Wesolowska-Andersen, C. Brorsson, A. L. Rajendrakumar, S. Hapca, S. Gan, A. Y. Dawed, L. A. Donnelly, R. McCrimmon, A. S. F. Doney, C. N. A. Palmer, V. Mohan, R. M. Anjana, A. T. Hattersley, J. M. Dennis and E. R. Pearson, "Heterogeneity in phenotype, disease progression and drug response in type 2 diabetes," *Nature Medicine,* vol. 28, pp. 982-988, 2022.

[15] M. S. Udler, J. Kim, M. v. Grotthuss, S. Bonàs-Guarch, J. B. Cole, J. Chiou, C. D. Anderson on behalf of METASTROKE and the ISGC, M. Boehnke, M. Laakso, G. Atzmon, B. Glaser, J. M. Mercader, K. Gaulton, J. Flannick, G. Getz and J. C. Florez, "Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis," *PLOS Medicine,* vol. 15, no. 9, p. e1002654, 2018.

[16] J. M. Dennis, B. M. Shields, W. E. Henley, A. G. Jones and A. T. Hattersley, "Clusters provide a better holistic view of type 2 diabetes than simple clinical features - author's reply," *The Lancet Diabetes & Endocrinology,* vol. 7, no. 9, p. 669, 2019.

[17] O. P. Zaharia, K. Strassburger, A. Strom, G. J. Bönhof, Y. Karusheva, S. Antoniou, K. Bódis, D. F. Markgraf, V. Burkart, K. Müssig, J.-H. Hwang, O. Asplund, L. Groop, E. Ahlqvist, J. Seissler, P. Nawroth, S. Kopf, S. M. Schmid, M. Stumvoll, A. F. H. Pfeiffer, S. Kabiscb, S. Tselmin, H. U. Häring, D. Ziegler, O. Kuss, J. Szendroedi, M. Roden and German Diabetes Study Group, "Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study," *Lancet Diabetes Endocrinology,* vol. 7, no. 9, pp. 684-694, 2019.

[18] J. L. Hill, "Bayesian nonparametric modelling for causal inference," *Bayesian Nonparametric,* vol. 20, no. 1, pp. 217-240, 2011.

[19] K. G. Young, E. H. McInnes, R. J. Massey, A. R. Kahkohska, S. J. Pilla, S. Raghaven, M. A. Stanislawski, D. K. Tobias, A. P. McGovern, A. Y. Dawed, A. G. Jones, E. R. Pearson, J. M. Dennis and ADA/EASD Precision Medicine in Diabetes Initiative, "Precision medicine in type 2 diabetes: a systematic review of treatment effect heterogeneity for GLP1-receptor agonists and SGLT2-inhibitors," *medRxiv,* 2023.

[20] R. V. Overgaard, C. L. Hertz, S. H. Ingwersen, A. Navarria and D. J. Drucker, "Levels of circulating semaglutide determine reductions in HbA1c and body weight in people with type 2 diabetes," *Cell Reports Medicine,* vol. 2, no. 9, p. 100387, 2021.

[21] P. W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association,* vol. 81, no. 396, pp. 945-960, 1985.

[22] I. M. Stratton, A. I. Adler, H. A. W. Neil, D. R. Matthews, S. E. Manley, C. A. Cull, D. Hadden, R. C. Turner and R. R. Holman, "Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study," *British Medical Journal,* vol. 321, no. 7258, pp. 405-412, 2000.

[23] R. R. Holman, S. K. Paul, M. A. Bethel, D. R. Matthews and H. A. W. Neil, "10-year follow-up of intensive glucose control in type 2 diabetes," *The New England Journal of Medicine,* vol. 359, no. 15, pp. 1577-1589, 2008.

[24] D. I. Patoulias, A. Katsimardou, M. S. Kalogirou, I. Zografou, M. Toumpourleka, K. P. Imprialos, K. Stavropoulos, I. Stergiou, C. Papadopoulos and M. Doumas, "Glucagon-like peptide-1 receptor agonists or sodium–glucose cotransporter-2 inhibitors as add-on therapy for patients with type 2 diabetes? a systematic review and meta-analysis of surrogate metabolic endpoints," *Diabetes & Metabolism,* vol. 46, no. 4, pp. 272-279, 2020.

[25] J. J. McMurray and N. Sattar, "Heart failure: now centre-stage in diabetes," *The Lancet Diabetes & Endocrinology,,* vol. 10, no. 10, pp. 689-691, 2022.

[26] A. Caron, "SparseBCF: sparse Bayesian causal forest for heterogeneous treatment," 2020.

[27] P. Cardoso, J. M. Dennis, J. Bowden, B. M. Shields and T. J. Beverley M, "Dirichlet process mixture models to estimate outcomes for individuals with missing predictor data: application to predict optimal type 2 diabetes therapy in electronic health record data," *medRxiv.*

[28] A. Wolf, D. Dedman, J. Campbell, H. Booth, D. Lunn, J. Chapman and P. Myles, "Data resource profile: clinical practice research datalink (CPRD) aurum," *International Journal of Epidemiology,* vol. 48, no. 6, pp. 1740-1740g, 2019.

[29] L. R. Rodgers, M. N. Weedon, W. E. Henley, A. T. Hattersley and B. M. Shields, "Cohort profile for the MASTERMIND study: using the clinical practice research datalink (CPRD) to investigate stratification of response to treatment in patients with type 2 diabetes," *BMJ Open,* vol. 7, p. e017989, 2017.

[30] National Institute for Health and Care Excellence, "Type 2 diabetes in adults: management," 2015. [Online]. Available: https://www.nice.org.uk/guidance/ng28. [Accessed 05 04 2023].

[31] A. Tsapas, I. Avgerinos, T. Karagiannis, K. Malandris, A. Manolopoulos, P. Andreadis, A. Liakos, D. R. Matthews and E. Bekiari, "Comparative effectiveness of glucose-lowering drugs for type 2 diabetes," *Annals of Internal Medicine,* vol. 173, no. 4, pp. 278-286, 2020.

[32] "Scottish Care Information-Diabetes Collaboration (SCI-Diabetes)," [Online]. Available: https://www.sci-diabetes.scot.nhs.uk.

[33] L. A. Inker, N. D. Eneanya, J. Coresh, H. Tighiouart, D. Wang, Y. Sang, D. C. Crews, A. Doria, M. M.

Estrella, M. Froissart, M. E. Grams, T. Greene, A. Grubb, V. Gudnason, O. M. Gutiérrez, R. Kalil, A. B. Karger, M. Mauer, G. Navis, R. G. Nelson, E. D. Poggio, R. Rodby, P. Rossing, A. D. Rule, E. Selvin, J. C. Seegmiller, M. G. Shlipak, V. E. Torres, W. Yang, S. H. Ballew, S. J. Couture, N. R. Powe, A. S. Levey and Chronic Kidney Disease Epidemiology Collaboration, "New creatinine- and cystatin c-based equations to estimate GFR without race," *The New England Journal of Medicine,* vol. 385, no. 19, pp. 1737-1749, 2021.

[34] H. A. Chipman, E. I. George and R. E. McCulloch, "BART: Bayesian additive regression trees," *Annals of Applied Statistics,* vol. 4, no. 1, pp. 266-298, 2010.

[35] A. Kapelner and J. Bleich, "bartMachine: Machine Learning with Bayesian Additive Regression Trees," *arXiv,* 2013.

[36] J. Bleich, A. Kapelner, E. I. George and S. T. Jensen, "Variable selection for BART: an application to gene regulation," *The Annals of Applied Statistics,* vol. 8, no. 3, pp. 1750-1781, 2014.

[37] J. Tibshirani, S. Athey, E. Sverdrup and S. Wager, *grf: Generalized Random Forests,* 2023.

[38] Y. Cui, M. R. Kosorok, E. Sverdrup, S. Wager and R. Zhu, "Estimating heterogeneous treatment effects with right-censored data via causal survival forests," *Journal of the Royal Statistical Society Series B: Statistical Methodology,* vol. 85, no. 2, pp. 179-211, 2023.

[39] F. E. Harrell, Regression modeling strategies, Springer International Publishing, 2016.

[40] M. Caliendo and S. Scheel-Kopeinig, "Some practical guidance for the implementation of propensity score matching," *Journal of Economic Surveys,* vol. 22, no. 1, pp. 31-72, 2008.

[41] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh and P. Brindle, "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2," *BMJ,* vol. 336, no. 7659, pp. 1475-1482, 2008.

[42] G. S. Collins, J. B. Reitsma, D. G. Altman and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," *Annals of Internal Medicine,* vol. 162, no. 1, pp. 55-63, 2015.

[43] J. M. Dennis, K. G. Young, A. P. McGovern, B. A. Mateen, S. J. Vollmer, M. D. Simpson, W. E. Henley, R. R. Holman, N. Sattar, E. R. Pearson, A. T. Hattersley, A. G. Jones and B. M. Shields, "Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in people with type 2 diabetes: a retrospective cohort study," *The Lancet Digital Health,* pp. 4(12): 873-883, 2022.

Tables

Table 1. Baseline clinical characteristics of patients initiating GLP-1 receptor agonists and SGLT2-inhibitors from the UK Clinical Practice Research Datalink.

Data are mean [SD] and number (%). SMD: Standardised mean difference (≥0.1 is a common metric for a meaningful imbalance between treatment groups).

	GLP-1 receptor agonists (n=28,081)		SGLT2 inhibitors (n=84,193)		SMD
		Missing (%)		Missing (%)	
Current age, years	57.7 [11.2]	-	58.4 [10.8]	-	0.064
Duration of diabetes, years	9.4 [6.4]	-	9.4 [6.6]	-	0.012
Year of drug start	2016.6 [2.2]	-	2017.4 [1.8]	-	0.390
Sex					
Male	14,960 (53.3%)	-	51,288 (60.9%)	-	0.155
Female	13,121 (46.7%)		32.905 (39.1%)		
Ethnicity					
White	24,063 (85.7%)		64,111 (76.1%)		
South Asian	2,144 (7.6%)	379	12,234 (14.5%)	1,786	0.255
Black	987 (3.5%)	(1.3%)	3,861 (4.6%)	(2.1%)	
Other	262 (0.9%)		1,316 (1.6%)		
Mixed	246 (0.9%)		885 (1.1%)		
SGLT2 inhibitor type					
Canagliflozin	-	-	14,965 (17.8%)	-	
Dapagliflozin	-	-	36,250 (43.1%)	-	
Empagliflozin	-	-	32,860 (39.0%)	-	
Ertugliflozin	-	-	137 (0.2%)	-	
GLP-1 receptor agonist type					-
Dulaglutide	10,337 (36.8%)	-	-	-	
Exenatide (short-acting)	1,367 (4.9%)	-	-	-	
Exenatide (long-acting)	2,377 (8.5%)	-	-	-	
Liraglutide	11,260 (40.1%)	-	-	-	
Lixisenatide	2,751 (9.8%)	-	-	-	
Index of multiple deprivation					
1 (Least deprived)	4,568 (16.3%)		14,143 (16.8%)		
2	4,925 (17.5%)	18	14,836 (17.6%)	47	0.033
3	5,393 (19.2%)	(0.1%)	16,194 (19.2%)	(0.1%)	
4	6,099 (21.7%)		18,841 (22.4%)		
5 (Most deprived)	7,078 (25.2%)		20,132 (23.9%)		
Smoking status					
Active	4,689 (16.7%)	1,244	14,265 (16,9%)	3,334	0.048
Ex-smoker	15,543 (55.4%)	(4.4%)	45,283 (53.8%)	(4.0%)	
Non-smoker	6,605 (23.5%)		21,311 (25.3%)		
Number of glucose-lowering drug classes ever prescribed					
2	2,886 (10.3%)	-	18,724 (22.2%)	-	0.420
3	6,772 (24.1%)		25,570 (30.4%)		
4	10,536 (37.5%)		25,217 (30.0%)		
5+	7,887 (28.1%)		14,682 (17.4%)		
Number of other current glucose--lowering drugs					
0	1,161 (5.8%)		5,155 (6.1%)		
1	9,948 (35.4%)	-	34,032 (40.4%)	-	0.137
2	12,422 (44.2%)		35,540 (42.2%)		
3	3,862 (13.8%)		9,187 (10.9%)		
4+	233 (0.8%)		279 (0.3%)		
Background therapy	24,075 (85.7%)	-	73,392 (87.2%)	-	0.042
Metformin	13,312 (47.4%)	-	30,165 (35.8%)	-	0.237
Sulfonylurea	4,595 (16.4%)	-	23,256 (27.6%)	-	0.274
DPP-4 inhibitor	4,019 (14.3%)	-	-	-	-

SGLT2 inhibitor	1,312 (4.7%)	-	2,374 (2.8%)	-	0.098
Thiazolidinedione	-	-	4,603 (5.5%)	-	-
GLP-1 receptor agonist					
Biomarkers					
HbA_{1c}, mmol/mol	78.6 [17.1]	5,795 (20.6%)	76.9 [16.9]	10,252 (12.2%)	0.101
BMI, kg/m^2	37.3 [7.2]	771 (2.7%)	33.7 [6.9]	3,234 (3.8%)	0.522
eGFR, mL/min per 1.3 m^2	92.0 [19.7]	68 (0.2%)	94.7 [15.5]	217 (0.3%)	0.152
HDL, cholesterol, mmol/L	1.1 [0.3]	1,333 (4.7%)	1.1 [0.3]	3,019 (3.6%)	0.083
Alanine transaminase, IU/L	35.1 [20.6]	1,801 (6.4%)	34.5 [20.2]	4,841 (5.7%)	0.027
Albumin, g/L	41.6 [3.9]	1,347 (4.8%)	42.0 [3.9]	3,667 (4.4%)	0.109
Bilirubin, µmol/L	9.1 [4.7]	1,187 (4.2%)	9.5 [5.0]	3,385 (4.0%)	0.084
Total cholesterol, mmol/L	4.4 [1.1]	167 (0.6%)	4.3 [1.1]	398 (0.5%)	0.039
Mean arterial blood pressure, mm Hg	96.1 [9.0]	82 (0.3%)	96.2 [9.0]	270 (0.3%)	0.016
Microvascular complications					
Nephropathy	730 (2.6%)	-	1,623 (1.9%)	-	0.045
Neuropathy	7,942 (28.3%)	-	20,161 (23.9%)	-	0.099
Retinopathy	10,540 (37.5%)	-	31,664 (37.6%)	-	0.002
Cardiovascular conditions					
Angina	3,224 (11.5%)	-	8,251 (9.8%)	-	0.055
Atherosclerotic cardiovascular disease*	6,285 (22.4%)	-	16,530 (19.6%)	-	0.068
Atrial fibrillation	1,737 (6.2%)	-	4,129 (4.9%)	-	0.056
Cardiac revascularisation	1,863 (6.6%)	-	5,632 (6.7%)	-	0.002
Heart failure	1,662 (5.9%)	-	3,654 (4.3%)	-	0.072
Hypertension	16,833 (59.9%)	-	46,550 (55.3%)	-	0.094
Ischaemic heart disease	4,181 (14.9%)	-	11,309 (13.4%)	-	0.042
Myocardial infarction	1,943 (6.9%)	-	5,655 (6.7%)	-	0.008
Peripheral arterial disease	1,703 (6.1%)	-	3,836 (4.6%)	-	0.067
Stroke	1,270 (4.5%)	-	3,422 (4.1%)	-	0.023
Transient ischaemic attack	766 (2.7%)	-	2,126 (2.5%)	-	0.013
Other conditions					
Chronic kidney disease	2,684 (9.6%)	-	2,962 (3.5%)	-	0.246
Chronic liver disease	3,754 (13.4%)	-	10,241 (12.2%)	-	0.036
QRISK2 10-year score	23.5 [13.5]	1,573 (5.6%)	23.5 [13.2]	6,215 (7.4%)	0.006
HbA_{1c} outcome					
HbA_{1c}, mmol/mol	66.3 [18.3]	15,397	64.2 [15.0]	40,025	0.126
Month of HbA_{1c} measure	8.9 [3.5]	(54.8%)	9.2 [3.5]	(47.5%)	0.071

*Atherosclerotic cardiovascular disease – composite of myocardial infarction, stroke, ischemic heart disease, peripheral arterial disease and revascularisation.

Table 2. Baseline clinical features included in the treatment selection algorithm after variable selection.

Features predictive of HbA1c outcome with SGLT2i	Features predictive of differential HbA1c outcome with GLP1-RA compared to SGLT2i
Current age, years	Current age, years
Duration of diabetes, years	Sex
Number of glucose-lowering drug classes ever prescribed	Number of other current glucose-lowering drugs
Number of other current glucose-lowering drugs	HbA_{1c}, mmol/mol
HbA_{1c}, mmol/mol	BMI, kg/m_2
eGFR, mL/min per 1.3 m^2	eGFR, mL/min per 1.3 m^2
Alanine transaminase, IU/L	Heart Failure
Peripheral arterial disease	Ischaemic heart disease
	Neuropathy
	Peripheral arterial disease
	Retinopathy

Figures

Fig. 1: Predicted conditional average treatment effects and model calibration

(A) Distribution of conditional average treatment effect (CATE) estimates for SGLT2-inhibitors vs. GLP-1 receptor agonists in the CPRD development cohort; negative values reflect a predicted HbA1c treatment benefit on SGLT2-inhibitors and positive values reflect a predicted treatment benefit on GLP-1 receptor agonists. (B) Calibration between average treatment effects (ATE) and predicted CATE estimates, by decile of predicted CATE in the development cohort. (C) Calibration of CATE estimates in the validation cohort. ATE estimates are adjusted for all the variables used in the treatment selection model (see Methods).

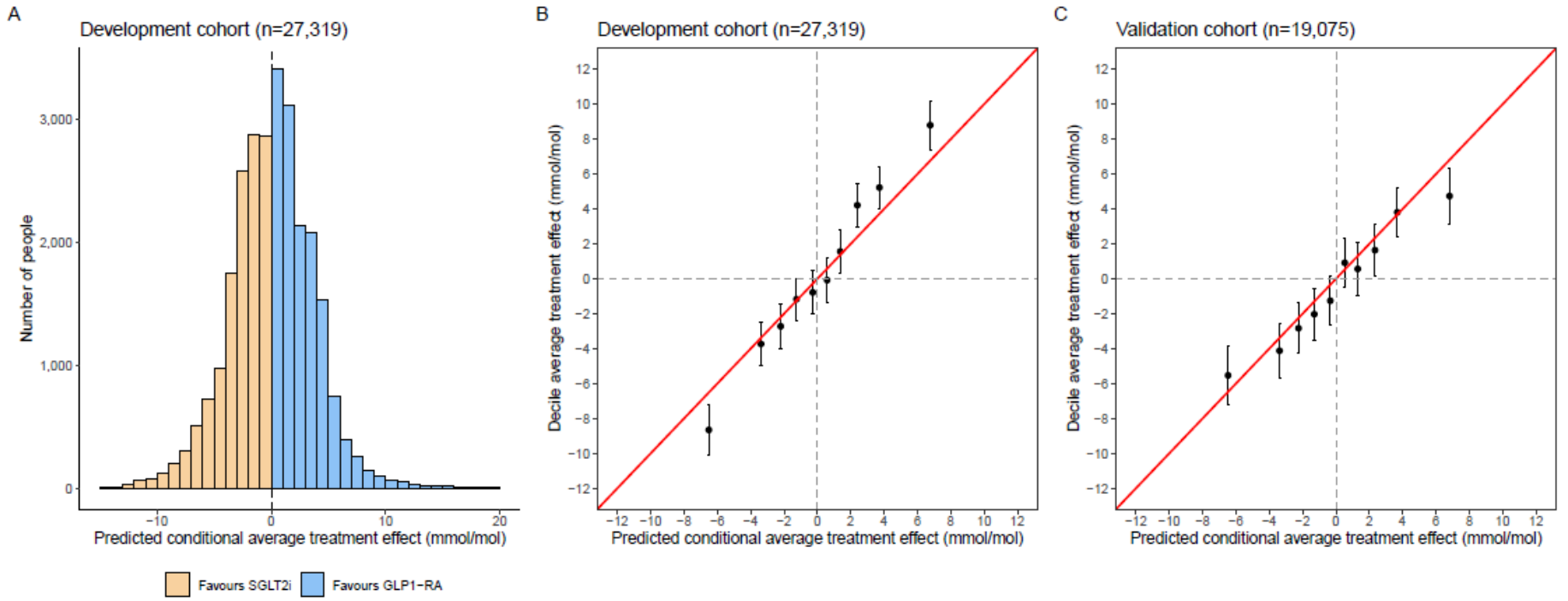


Fig. 2: Distributions of major clinical characteristics predicting differential HbA1c outcome with SGLT2i and GLP1-RA

Distributions of key differential clinical characteristics in the combined development and validation cohorts (n=46,394 with complete predictor data) for subgroups defined by predicted HbA_{1c} outcome differences: SGLT2i benefit ≥5 mmol/mol, 3–5 mmol/mol and 0–3 mmol/mol, GLP1-RA benefit ≥5 mmol/mol, 3–5 mmol/mol and 0–3 mmol/mol.

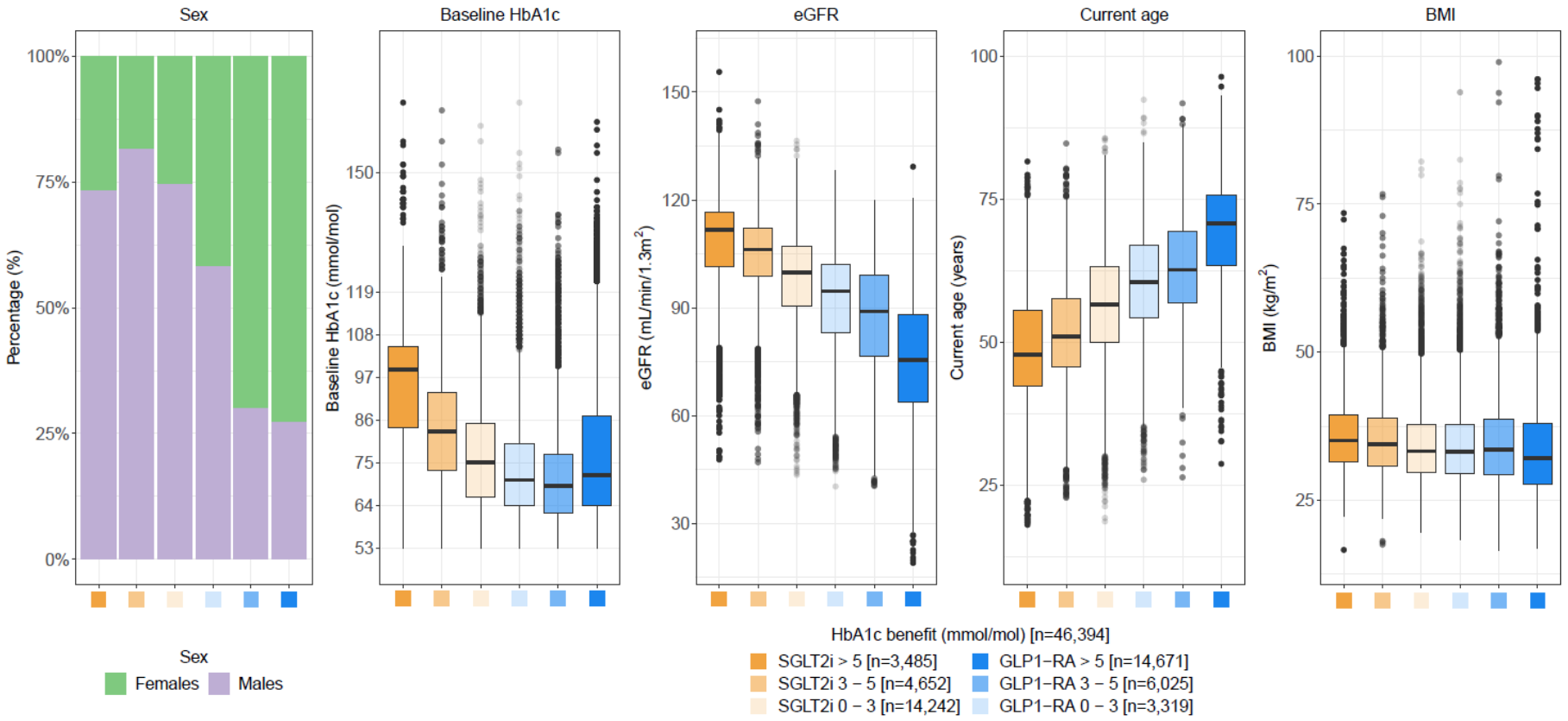
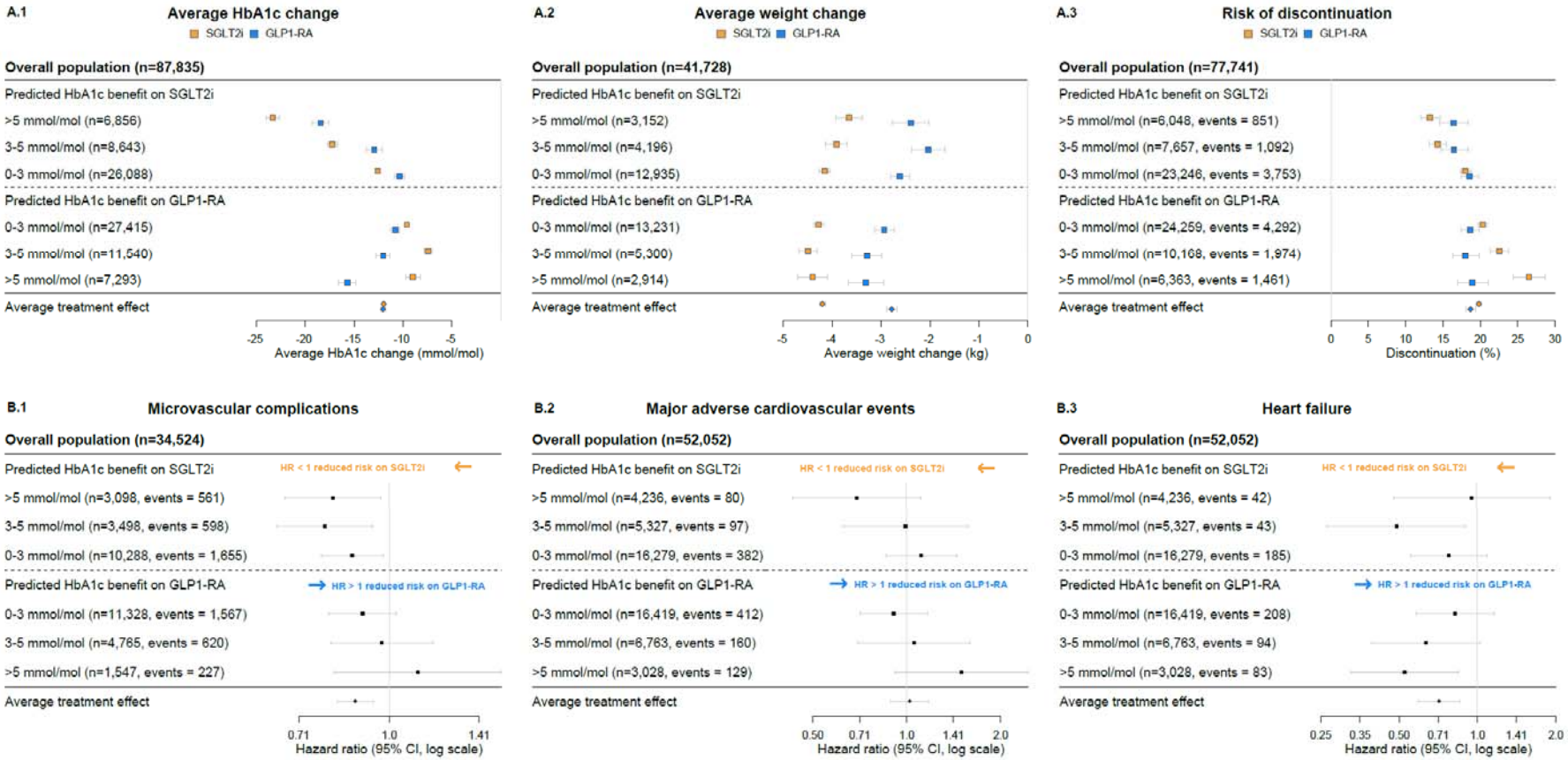


Fig. 3: Differences in short-term and long-term clinical outcomes with SGLT2i and GLP1-RA for subgroups defined by predicted HbA1c response differences

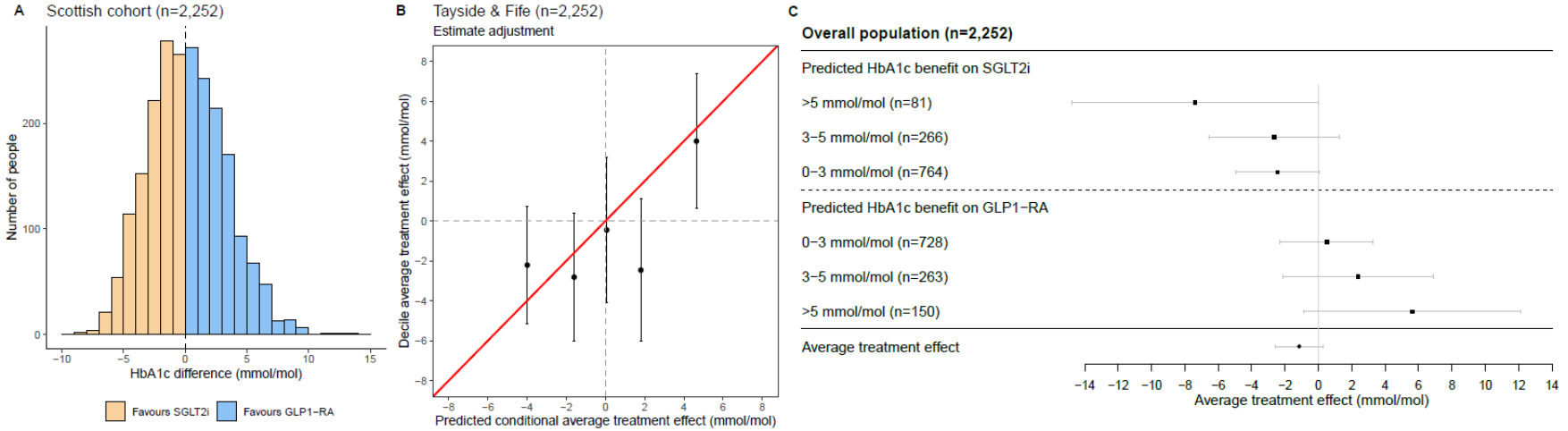
(A.1) 12-month HbA1c change from baseline. (A.2) 12-month weight change. (A.3) 6-month risk of discontinuation. (B.1) Hazard ratios for 5-year risk of new-onset microvascular complications (retinopathy, nephropathy or neuropathy). (B.2) Hazard ratios for 5-year relative risk of major adverse cardiovascular events (MACE). (B.3) Hazard ratios for 5-year risk of heart failure. Hazard ratios represent the relative risk for those treated with GLP1-RA in comparison to SGLT2i therapy, with a value under 1 favouring SGLT2i therapy. Data underlying the Figure are reported in sTable 3.



Extended Figures

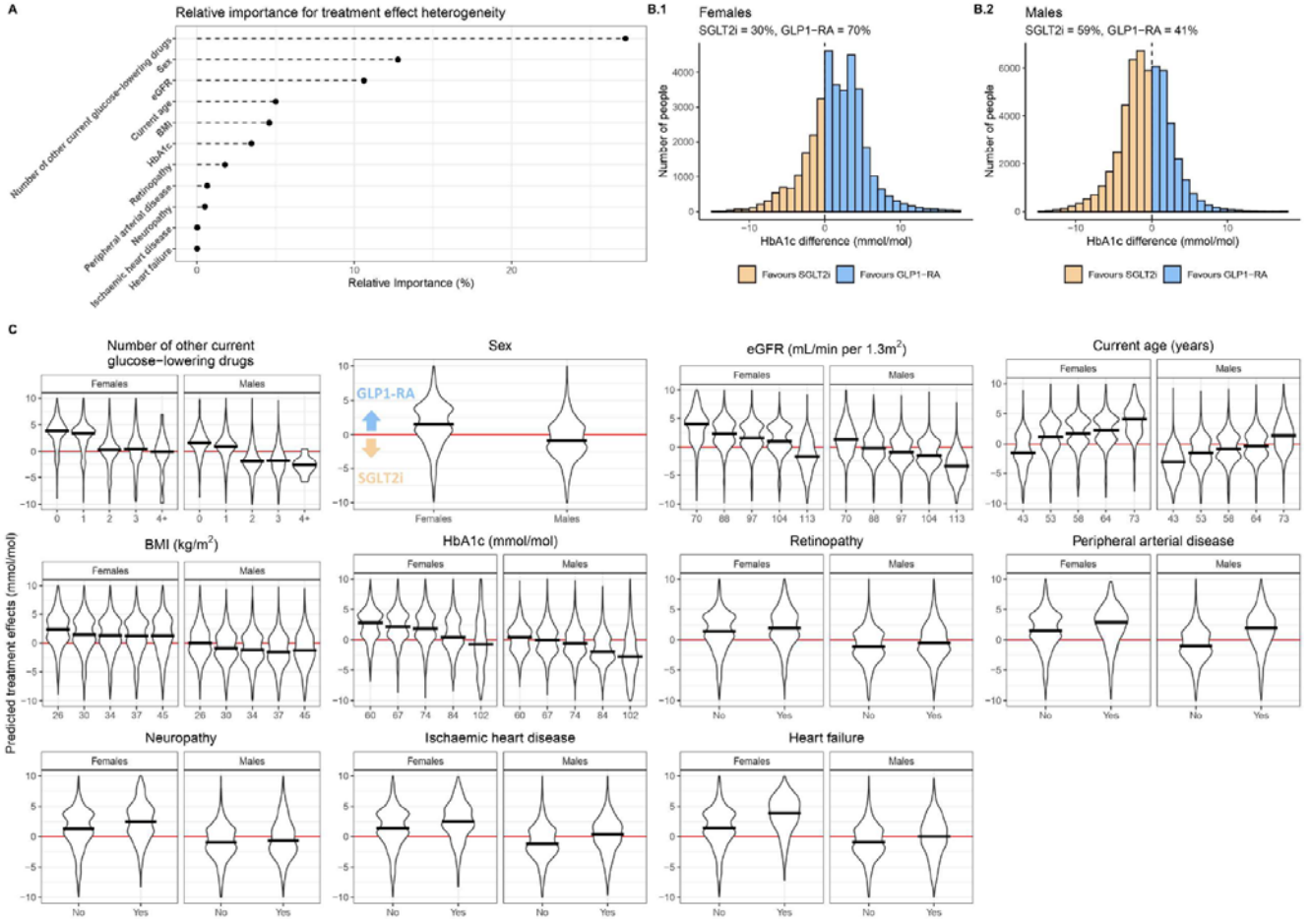
Extended Fig. 1: External validation in Tayside & Fife, Scotland

(A) Distribution of conditional average treatment effect (CATE) estimates for SGLT2i vs. GLP1-RA; negative values reflect a predicted glucose-lowering treatment benefit on SGLT2-inhibitors and positive values reflect a predicted treatment benefit on GLP-1 receptor agonists. (B) Calibration between average treatment effects (ATE) and predicted CATE estimates, by quintile of predicted CATE. (C) ATE estimates within subgroups defined by clinically meaningful CATE thresholds (SGLT2i benefit >5, 3-5 and 0-3 mmol/mol, GLP1-RA benefit >5, 3-5 and 0-3 mmol/mol).



Extended Fig. 2: Model interpretability plots

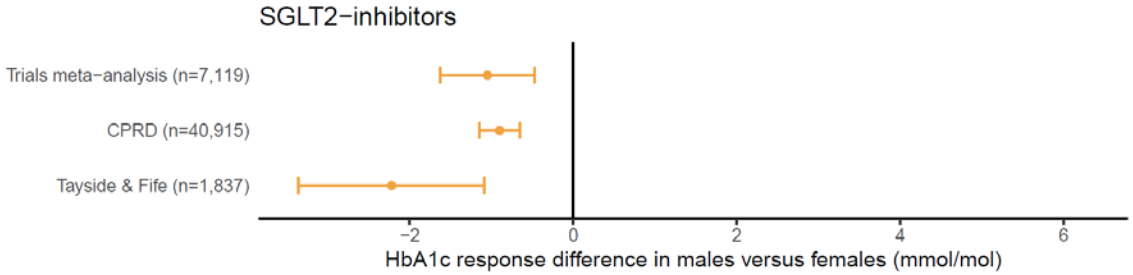
(A) Relative variable importance for clinical features predicting differential treatment effects (best linear projection of BCF model; see Methods). (B) Distribution of conditional average treatment effect (CATE) estimates for SGLT2i vs. GLP1-RA, by sex. (C) Predicted treatment effects for all differential clinical features, with individuals stratified into quintiles for continuous variables, and black lines corresponding to the median of the stratified group. All estimates are for the overall study population, n= 46,394.



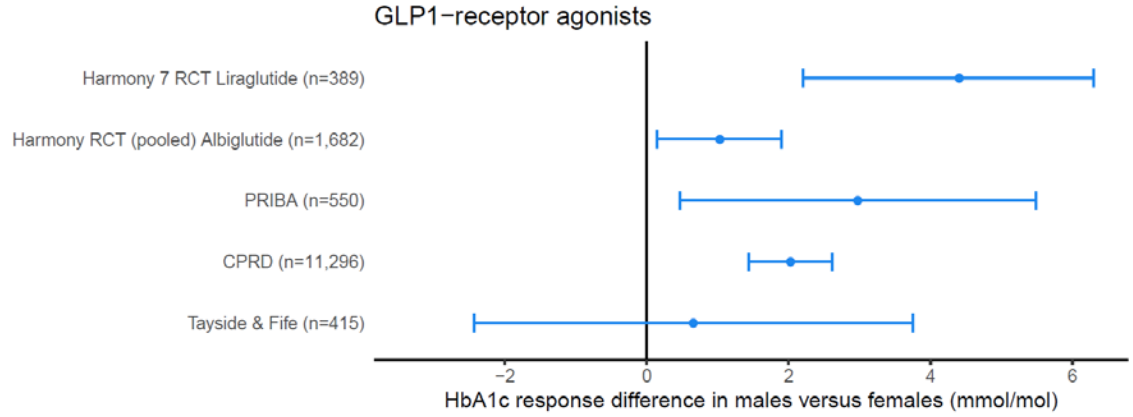
Extended Fig. 3: Differences in HbA1c outcome by sex, in randomized clinical trial and observational datasets.

All estimates are adjusted for baseline HbA1c. Estimates lower than zero represent a greater HbA1c reduction in males compared to females. Bars represent 95% confidence intervals.

A) SGLT2-inhibitors. Point estimates for the trials meta-analysis and CPRD are reproduced from Dennis *et al.* [43].



B) GLP1-receptor agonists.



Extended Fig. 4: Differential HbA1c treatment effects in individuals of white and non-white ethnicity

Adjusted ATE estimates within subgroups defined by clinically meaningful CATE thresholds (SGLT2i benefit >5, 3-5 and 0-3 mmol/mol, GLP1-RA benefit >5, 3-5 and 0-3 mmol/mol). Negative values reflect a predicted glucose-lowering treatment benefit with SGLT2i, and positive values reflect a predicted treatment benefit with GLP1-RA. The non-white subgroup is a composite of major UK non-white self-reported ethnicity groups: Black, South Asian, Mixed and Other. Individuals without a recorded ethnicity were excluded (n=932)