A phenomapping-derived tool to individualize canagliflozin's effect on cardiovascular risk in type 2 diabetes

Online-only supplement

Evangelos K Oikonomou MD, DPhil[1], Marc A Suchard MD, PhD[2,3],
Darren K. McGuire, MD, MHSc[4], Rohan Khera MD, MS[1,5]

[1] Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

[2] Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA, USA

[3] Departments of Computational Medicine and Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

[4] University of Texas Southwestern Medical Center and Parkland Health and Hospital System, Dallas, TX USA

[5] Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA

Brief title: Personalizing SGLT2 inhibitor therapy

Address for correspondence:

Rohan Khera, MD, MS

1 Church St, Suite 200, New Haven, CT 06510

203-764-5885; rohan.khera@yale.edu; @rohan_khera

Supplementary Materials

Outcome definitions

To ensure consistency with the original CANVAS program trials (1), we complied with the outcome definitions described in the trial protocols. We defined cardiovascular mortality as a death event attributed to any cardiac or vascular disorder, including central nervous system vascular disorders, pulmonary embolism and pulmonary edema. Deaths of unknown cause were also included in this group, as per the original CANVAS trials protocols. Any death not classified as cardiovascular was defined as non-cardiovascular.

Medication definitions

Medications used at baseline were extracted from the baseline, run-in visit records based on the ATC (Anatomical Therapeutic Chemical Classification System) Index 2021, a system of alphanumeric codes developed by the World Health Organization (WHO) for the classification of drugs and other medical products. The codes used for the definition of each medication group are summarized below:

Diuretics: ‘C03’, 'C01AA', 'C02L', 'C07B', 'C07C', 'C07D', 'C08G', 'C09BA', 'C09DA'

Beta blockers: ‘C07’, 'C09BX', 'C09DX'

RAAS inhibitors: 'C09', 'C10BX'

Calcium channel blockers: 'C08', 'C04AE', 'C08G', 'C09BB', 'C07FB', 'C10BX'

Statins: 'C10AA', 'C10BA02', 'C10BA03', 'C10BA04', 'C10BA05', 'C10BA06', 'C10BA07', 'C10BA08', 'C10BA09', 'C10BX'

Antiplatelets: 'B01AC'

Anticoagulants: 'B01AA', 'B01AB', 'B01AE', 'B01AF', 'B01AX'

Insulin: 'A10A'

Insulin (short-acting): 'A10AB', 'A10AF'

Insulin (intermediate- or long-acting): 'A10AC', 'A10AE', 'A10AD'

Biguanides: 'A10BA', 'A10BD01', 'A10BD02', 'A10BD03', 'A10BD05', 'A10BD07', 'A10BD08', 'A10BD10', 'A10BD11', 'A10BD13', 'A10BD14', 'A10BD15', 'A10BD16', 'A10BD17', 'A10BD18', 'A10BD20', 'A10BD22', 'A10BD23', 'A10BD25', 'A10BD26'

Sulfonylureas: 'A10BB', 'A10BD01', 'A10BD02', 'A10BD04', 'A10BD06'

Thiazolidinediones: 'A10BG', 'A10BD04', 'A10BD05', 'A10BD06', 'A10BD09', 'A10BD12', 'A10BD26'

Dipeptidyl-peptidase IV inhibitors: 'A10BH', 'A10BD07', 'A10BD08', 'A10BD09', 'A10BD10', 'A10BD11', 'A10BD12', 'A10BD13', 'A10BD18', 'A10BD19', 'A10BD21', 'A10BD22', 'A10BD24', 'A10BD25'

Glucagon-like peptide 1 agonists: 'A10BJ'

Other antihyperglycemic medications: 'A10BX'

Missing data imputation

We imputed missing variables using chained random forests with predictive mean matching deployed within the *missRanger* package in R. This approach enables imputation of mixed-type datasets while avoiding imputed values not already present in the original data, while also raising the variance in the resulting conditional distributions to a realistic level.

Dissimilarity distance calculation using Gower's method

As per our previous work (2), we first computed a dissimilarity index that classified individuals based on their detailed clinical characteristics according to the Gower distance. For numeric variables, Gower's distance calculates the absolute value of the difference divided by the range. For non-numeric elements (categorical variables) the method assigns "1" if the values are identical and "0" if they are not. Gower's distance is ultimately calculated as the average of these terms (3).

Visualization using uniform manifold approximation and projection

To visualize phenotypic variation in the CANVAS population and neighborhoods, we used a dimensionality reduction method with uniform manifold approximation and projection (UMAP) (4). UMAP constructs a high-dimensional graph and then optimizes a low-dimensional graph to be as structurally similar as possible. UMAP aims to maintain a balance between the local and global structure of the data by decreasing the likelihood of connection as the outwards radius around each datapoint increases, thus maintaining the local architecture while ensuring that each point is connected to at least its closest neighbor and ensuring a global representation (4). Two visualization parameters – the number of neighbors and the minimum distance – were set at 200 and 0.6, respectively, both chosen to optimize how well the UMAP learned the manifold structure of the data and preserved the broad topological structure of our dataset while maintaining the visualization. These values only affect the visualization but not the structure of the topological space.

Training of the extreme gradient boosting algorithm

Data preprocessing: For the training of the extreme gradient boosting algorithm, categorical variables with more than two groups were one-hot encoded (0: no, 1: yes) and we removed groups with near-zero variance (frequency cut-off for ratio of most common value to the second most common value of 95 to 5).

Machine learning task creation: An extreme gradient boosting algorithm was trained using the CANVAS data using a tree or linear gradient booster. In the tree gradient booster, a tree is grown one after another and attempts to reduce the error rate in subsequent iterations. At each level the next tree is built by giving a higher weight to points with the highest residuals (actual-predicted). We set our problem as a regression task, using root mean squared error as our metric to evaluate our model's accuracy in the validation data.

Hyperparameter tuning: We performed a total of 100 iterations with hyperparameter tuning validated using five-fold cross-validation. For the hyperparameter tuning, a grid was created using the following parameters: "gradient booster": tree or linear; "maximal depth of the tree": 3 to 10; "minimum number of instances required in a child node": 1 to 10; "number of samples supplied to a tree": 0.5 to 1; "number of features supplied to a tree": 0.5 to 1; " regularization parameter gamma" (penalizes large coefficients at higher values to prevent overfitting): 0 to 100; "L2 regularization parameter alpha" (equivalent to ridge regression): 0 to 1; and the "L1 regularization parameter lambda" (equivalent to lasso regression with shrinkage and feature selection): 0 to 1. Following a total of 100 iterations, the optimal parameters were identified as a tree gradient booster with "maximal depth of the tree": 10; "minimum number of instances required in a child node": 1.65; "number of samples supplied to a tree": 0.569; "number of

features supplied to a tree": 0.552; "gamma": 3.46; "alpha": 0.139; and "lambda": 0.334. Using the selected hyperparameters, we subsequently trained our model for a maximum of 500 iterations, with an early stopping function triggered if the performance in the validation set did not improve after 20 rounds (in our case this was reached after 194 iterations).

Feature importance evaluation and selection: We evaluated the importance of our features in the final model using the model-agnostic SHAP (Shapley Additive exPlanations) values, based on the theoretical concept of Shapley values as described in cooperative game theory. SHAP values shed light into the "black-box" nature of machine learning algorithms by interpreting the contribution of each input variable to the final prediction. SHAP values measure the impact of each variable considering the interaction with other variables. We visualized these using a SHAP summary plot, in which the vertical axis represents the variables in descending order of importance and the horizontal axis indicates the change in prediction (with wider bars along the horizontal axis associated with higher feature importance). The gradient color denotes the original value for that variable (for instance, for binary variables such as hypertension or diabetes, it only takes two colors, whereas for continuous variables, it contains the whole spectrum). In these plots. each point represents an individual from the original training set. To create an easy-to-use clinical model, we selected the top 15 features (feature importance of 0.01 or higher) and retrained our model, using five-fold cross-validation in CANVAS dataset, following the same methodology as the one reviewed above. The final algorithm was named INSIGHT (INdividualized cardiovaScular rIsk reduction with sGlt2 inHibiTors).

Supplemental tables

Table S1. Variables included in the pre-processing, training and validation stages.

Variable name	Available in >80% of CANVAS	Included in model training	Included in final model
Age	✓	✓	
Sex	✓	✓	✓
Race	✓	✓	✓
Ethnicity	✓	✓	✓
Dyslipidemia	✓	✓	✓
Hypertension	✓	✓	✓
Retinopathy	✓	✓	✓
Nephropathy	✓		
Neuropathy	✓	✓	✓
Coronary artery disease	✓	✓	✓
Cerebrovascular disease	✓		
Peripheral arterial disease	✓		
Heart failure	✓	✓	✓
History of amputation	✓		
History of coronary revascularization	✓		
Diuretic use	✓		
Beta blocker use	✓		
Renin-angiotensin-aldosterone blocker use	✓		
Calcium channel blocker use	✓		
Antiplatelet use	✓		
Anticoagulant use	✓		
Statin use	✓		
Insulin use (any)	✓		
Insulin use (intermediate-long)	✓		
Insulin use (short-acting)	✓		
Biguanide use	✓		
Thiazolidinedione use	✓		
Dipeptidyl-peptidase IV inhibitor use	✓		
Glucagon-like peptide 1 receptor agonist use	✓		
Other antihyperglycemic agent use	✓		
Smoking history	✓	✓	✓
Weight	✓		
Pulse	✓	✓	
Systolic blood pressure	✓	✓	
Diastolic blood pressure	✓	✓	
Height	✓		
Body mass index	✓	✓	✓
Albumin, serum	✓	✓	
Albumin, urine	✓		
Albumin/creatinine ratio, urine	✓	✓	
Alkaline phosphatase, serum	✓	✓	
Alanine aminotransferase, serum	✓		

Aspartate aminotransferase, serum	✓	✓	
Basophil count, blood	✓		
Basophil percentage, blood	✓		
Bicarbonate, serum	✓	✓	
Bilirubin total, serum	✓	✓	
Bilirubin, urine			
Blood, urine			
Blood urea nitrogen, urine	✓	✓	
Calcium, serum	✓	✓	
Total cholesterol, serum	✓		
Creatine kinase, serum	✓	✓	✓
Chloride, serum	✓	✓	
C-peptide, serum			
Creatinine, serum	✓		
Creatinine, urine	✓		
Eosinophil count, blood	✓	✓	
Eosinophil percentage, blood	✓		
Glomerular filtration rate	✓	✓	✓
Gamma-glutamyl transferase, serum	✓	✓	
Glucose, serum	✓		
Glycated hemoglobin A1c	✓	✓	
High-density lipoprotein cholesterol level, serum	✓	✓	✓
Insulin, serum			
Potassium, serum	✓		
Ketones, urine			
Lactate dehydrogenase, serum	✓	✓	
Low-density lipoprotein, serum	✓		
LDL-C to HDL-C ratio, serum	✓	✓	
Leukocyte esterase, urine			
Lymphocyte count, blood	✓	✓	
Lymphocyte percentage, blood	✓		
Magnesium, serum	✓	✓	
Monocyte count, blood	✓	✓	
Monocyte percentage, blood	✓		
Neutrophil count, blood	✓		
Neutrophil percentage, blood	✓		
Neutrophil-to-lymphocyte ratio, blood	✓	✓	
Nitrite, urine			
pH, urine	✓		
Phosphate, serum	✓	✓	
Platelet count, blood	✓	✓	
Proinsulin (pmol, mIU)			
Proinsulin (pmol, L)			
Protein total, serum	✓	✓	
Protein, urine			
Total right blood cell count, blood	✓		
Right blood cell morphology, blood		✓	
Sodium, serum	✓	✓	

Specific gravity, urine	✓		
Triglyceride, serum	✓	✓	
Uric acid, serum	✓	✓	
Urobilinogen, urine			
Polychromasia, blood			
Total white blood cell count, blood	✓	✓	
Atypical lymphocytes, blood			
Atypical lymphocytes percentage, blood			
Follicle-stimulating hormone, serum			
hCG, urine			
Thyroid-stimulating hormone, serum			
Anisocytosis, blood			
Elliptocytes, blood			
HpoRBC, blood			
Microcytosis, blood			
Rouleaux, blood			
Macrocytosis, blood			
Basophilic stippling, blood			
Neutrophil band count, blood			
Neutrophil band form to leukocyte ratio, blood			
Myelocyte count, blood			
Myelocyte count, percentage			
Bacteria, urine			
Bilirubin indirect, serum			
Epithelial cells, urine			
Poikilocytosis, blood			
QTc (Friderichia)	✓		
Heart rate	✓		
QT	✓		
QRS	✓		
PR	✓		
RR	✓		
QTc (Bazett's)	✓		

Table S2. Baseline characteristics of the CANVAS and CANVAS-R populations.

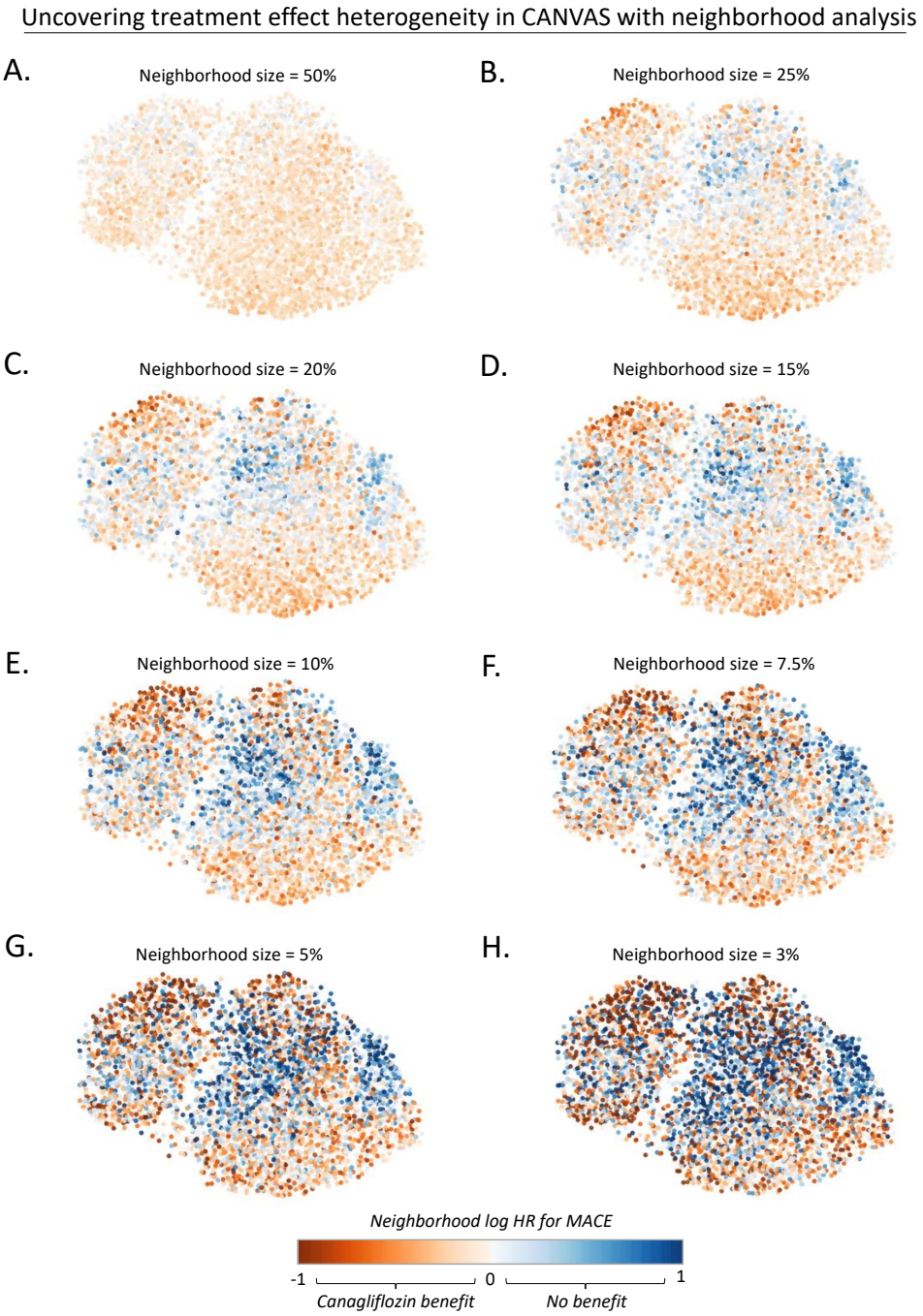
	CANVAS (n=4327)		CANVAS-R (n=5808)	
	Canagliflozin	**Placebo**	**Canagliflozin**	**Placebo**
Total number	2886	1441	2904	2904
Age (years)	61.0±8.2	60.8±8.0	62.3±8.5	62.4±8.4
Male (sex)	1905 (66.0)	955 (66.3)	1851 (63.7)	1793 (61.7)
Hispanic/Latino ethnicity	266 (9.2)	149 (10.3)	604 (20.8)	586 (20.2)
Race				
Native American	2 (0.1)	3 (0.2)	14 (0.5)	19 (0.7)
Asian	533 (18.5)	262 (18.2)	244 (8.4)	245 (8.4)
Black	70 (2.4)	35 (2.4)	105 (3.6)	125 (4.3)
Native Hawiian or Pacific Islander	5 (0.2)	1 (0.1)	8 (0.3)	9 (0.3)
Other	4139 (4.8)	66 (4.6)	132 (4.5)	135 (4.3)
White	2114 (73.3)	1063 (73.8)	2391 (82.3)	2371 (81.6)
Dyslipidemia	1980 (68.6)	969 (67.2)	2061 (71.0)	2100 (72.3)
Hypertension	2520 (87.3)	1271 (88.2)	2657 (91.5)	2660 (91.6)
Retinopathy	630 (21.8)	303 (21.0)	683 (23.5)	709 (24.4)
Nephropathy	449 (15.6)	228 (15.8)	575 (19.8)	576 (19.8)
Coronary artery disease	1462 (50.7)	728 (505.)	1355 (46.7)	1361 (46.9)
Medications				
Diuretic use	1275 (44.2)	721 (50.0)	Not provided	Not provided
Beta blocker use	1712 (59.3)	878 (60.9)	Not provided	Not provided
RAASi	2492 (86.3)	1269 (88.1)	Not provided	Not provided
Calcium channel blocker use	1180 (40.9)	643 (44.6)	Not provided	Not provided
Antiplatelet use	2157 (74.7)	1089 (75.6)	Not provided	Not provided
Statin use	2282 (79.1)	1126 (78.1)	Not provided	Not provided
Insulin use	2067 (71.6)	1103 (76.5)	Not provided	Not provided
Glycated hemoglobin	8.2±0.9	8.2±0.9	8.3±0.9	8.3±1.0
Summary statistics are presented as numbers (percentages, %) or mean ± standard deviation.				

Table S3. Risk of serious adverse events with canagliflozin therapy among different groups of predicted atherosclerotic benefit using INSIGHT.

	Predicted high responders*	Predicted low responders*	P value for interaction
	Total numbers, or Hazard Ratio (95% CI)		
CANVAS trial			
Total numbers	N=1224	N=3103	-
Risk of serious adverse event for canagliflozin versus placebo	0.91 (0.77-1.07)	1.00 (0.89-1.11)	0.36
Risk of any serious adverse event resulting in discontinuation of therapy for canagliflozin versus placebo	0.89 (0.69-1.15)	1.17 (0.97-1.41)	0.09
CANVAS-Renal trial			
Total numbers	N=1702	N=4106	-
Risk of serious adverse event for canagliflozin versus placebo	0.81 (0.68-0.98)	0.90 (0.80-1.02)	0.33
Risk of any serious adverse event resulting in discontinuation of therapy for canagliflozin versus placebo	0.64 (0.47-0.88)	0.88 (0.73-1.07)	0.09

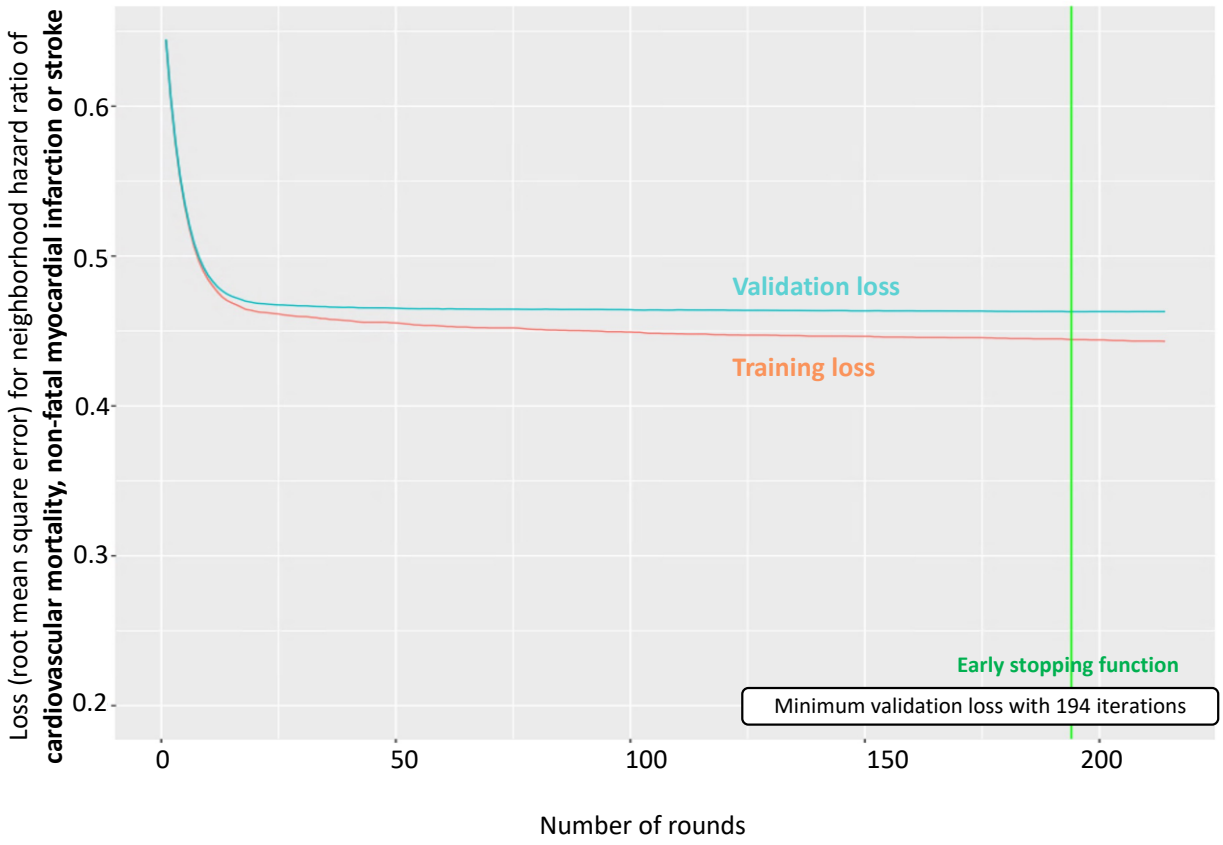
Hazard ratios are derived from Cox regression models with time to first serious adverse event (or serious adverse event requiring discontinuation of therapy) as the dependent outcome of interest and treatment arm as the independent variable further adjusted for age at randomization and sex. *High responders were defined as those individuals with a predicted benefit greater than half a standard deviation lower than the average predicted response (log HR) as defined in the methods.

CANVAS: Canagliflozin Cardiovascular Assessment Study; CI: confidence interval; INSIGHT: INdividualized cardiovaScular rIsk reduction with sGlt2 inHibiTors.

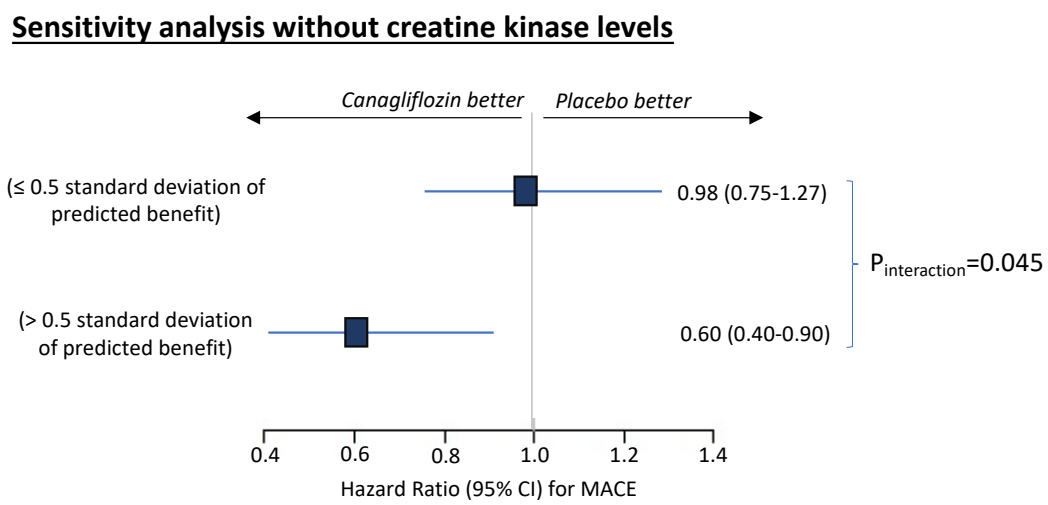


Supplemental Figure S1. Neighborhood analyses to uncover treatment effect heterogeneity. Using the manifold representation of all 4327 participants included in CANVAS, we defined 4327 neighborhoods of variable size (one around each study participant) and in each one,

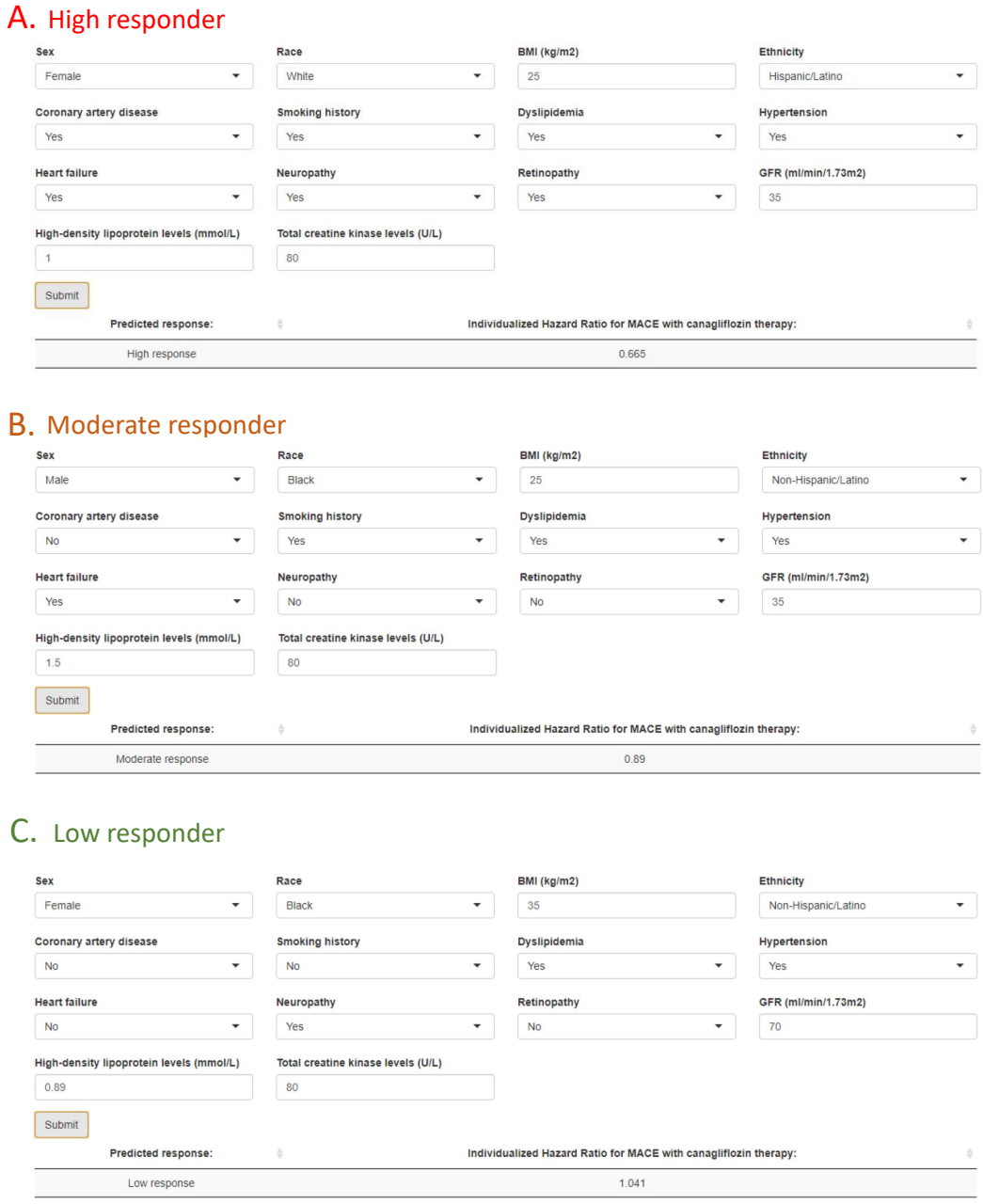
calculated age- and sex-adjusted risk estimates of major adverse cardiac events (MACE) with canagliflozin therapy versus placebo. Moving from larger neighborhoods (e.g. 50%, 25%, 20%, 15% of all participants (**A-D**)) to smaller neighborhoods (e.g. 10, 7.5%, 5%, 3% of all participants (**E-H**)) uncovered treatment effect heterogeneity that was non-uniformly distributed in the topological space. Horizontal and vertical axes which represent the first and second uniform manifold approximation and projection dimensions have been omitted.



Supplemental Figure S2. Learning curve for a 15-variable tool predicting the individualized benefit of canagliflozin for the primary MACE outcome in CANVAS. Panel depicts the loss function (root mean square error) with five-fold cross-validation across several rounds of training in the CANVAS trial with the optimal validation loss reached after a total of 194 iterations (rounds). INSIGHT: INdividualized cardiovaScular rIsk reduction with sGlt2 inHibiTors. MACE: major adverse cardiovascular events; RMSE: root mean square error.



Supplemental Figure S3. Observed risk across high and low predicted canagliflozin benefit groups based on the INSIGHT tool, after excluding circulating creatine kinase levels. hospitalization in CANVAS-R. CI: confidence interval; INSIGHT: INdividualized cardiovaScular rIsk reduction with sGlt2 inHibiTors; MACE: major adverse cardiac events.



Supplemental Figure S4. An online browser-accessible version of the INSIGHT decision support tool. A browser-accessible version of the INSIGHT decision support tool is available which enables its prospective and external application. The figure demonstrates three examples of patients in whom the use of canagliflozin is associated with a high (**A**), intermediate (**B**), and

low (**C**) predicted reduction in their respective ASCVD risk. INSIGHT: INdividualized cardiovaScular rIsk reduction with sGlt2 inHibiTors.

Supplementary references

1. Neal B, Perkovic V, Mahaffey KW et al. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. N Engl J Med 2017;377:644-657.
2. Oikonomou EK, Van Dijk D, Parise H et al. A phenomapping-derived tool to personalize the selection of anatomical vs. functional testing in evaluating chest pain (ASSIST). Eur Heart J 2021;doi: 10.1093/eurheartj/ehab223.
3. Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971:857-871.
4. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018:arXiv:1802.03426.