Cluster Analysis of Cardiovascular Phenotypes and SGLT2 Inhibition in Patients With Type 2 Diabetes and Established Cardiovascular Disease

Narrative Summary

Diabetes is the 9th leading cause of death worldwide. Although it has been studied for several decades, it remains a strong risk factor for illnesses affecting the heart and kidneys, even when treated. Evidence suggests that this may be because several unknown types (or “compositions”) of the disease exist. As a result, our study proposes using machine learning on people with diabetes to uncover hidden compositions of the illness, which may have different prognoses and reactions to treatment. These findings will help doctors provide personalized care by identifying patients who may need increased antidiabetic care, and likewise, patients who may benefit most from certain treatments.

Research Proposal

Despite significant advances in the design of therapies for people with type 2 diabetes mellitus (T2DM), the disease continues to portend high rates of morbidity and mortality, even when traditional cardiovascular risk factors are well controlled. Like disease states such as heart failure and atherosclerosis, there is significant evidence that this may be because numerous pathophysiological phenotypes of the disease exist. Although diabetes has historically been diagnosed into two classes (i.e., type I and type II), these data suggest that our binary approach to treatment may not be sufficient for risk reduction. As a result, unsupervised learning (e.g., machine learning) algorithms such as latent class analyses have proliferated in the clinical literature to improve personalized care.

As an alternative to traditional subgroup analysis, latent class analyses are a data agnostic method for elucidating distinct clinical phenotypes and their associated response to treatment. This method has been utilized successfully in a variety of medical disciplines, including heart failure with reduced or preserved ejection fraction. More recently, cluster analysis was utilized retrospectively in the Empagliflozin, Cardiovascular Outcome, and Mortality in T2DM (EMPA-REG outcomes) trial, identifying a nonsignificant trend of greater benefit in one of three phenotypes as characterized by lower rates of heart failure hospitalization in young people with...
low comorbidity burden. Despite these findings, some uncertainty persists as relatively few patients experienced these endpoints.

Thus, we propose conducting a pooled latent class analysis in patients from the CANVAS (Canagliflozin Cardiovascular Assessment Study) and CREDENCE (Canagliflozin and Renal Events in Diabetes with Established Nephropathy Clinical Evaluation) trials to assess the differential effects of SGLT2 inhibition on cardiovascular and renal outcomes in patients with T2DM and established cardiovascular disease. Such an analysis could aid in the identification of phenotypes (or “clusters”) of people with T2DM who may respond most beneficially to SGLT2 inhibition. This could help researchers identify novel patient groups for investigating new glycemic and cardiovascular treatments, potentially saving valuable resources, and allowing health care providers to offer greater access to care.

**Scientific Abstract**

**Background** In the CANVAS and CREDENCE trials, canagliflozin was associated with 33 and 39% reduction in the incidence of heart failure hospitalization, respectively. However, whether differential treatment effects exist with canagliflozin remains unclear.

**Objective** As a result, this study aims to utilize latent class analysis to identify distinct clinical phenotypes in subjects with T2D and cardiovascular disease to elucidate potential differences in treatment effects across clinical phenotypes.

**Study Design** Latent class analysis will be utilized to identify unobserved (or “latent”) subclasses of individuals with T2D and cardiovascular disease. Cox proportional hazard regression models with interaction terms will be utilized to assess whether cluster membership is associated with a differential response to canagliflozin.

**Participants** The population of interest for the proposed analysis encompasses the entire patient populations enrolled in the CANVAS and CREDENCE trials. In analyses where differential survival according to cluster membership is evaluated, patients will be subdivided into their respective trial arms (i.e., canagliflozin, placebo).

**Main Outcomes** The co-primary endpoints for the proposed analysis are time to heart failure hospitalization, and the composite of heart failure hospitalization or cardiovascular death. Secondary efficacy endpoints include cardiovascular death, nonfatal myocardial infarction, nonfatal stroke, cause specific mortality, all-cause mortality, regression of albuminuria, and two CKD progression-related composite outcomes: the composite outcome of death from renal causes,
sustained 40% reduction in eGFR or need for renal replacement therapy, and the composite outcome of doubling serum creatinine, end-stage renal disease, or death from renal causes. Secondary safety endpoints include any adverse event, any serious adverse event, serious adverse event related to canagliflozin, amputation, diabetic ketoacidosis, acute kidney injury, and hyperkalemia.

**Statistical Analysis** We will assess proportional hazards assumptions, and survival models will be adjusted for baseline clinical characteristics. In a sensitivity analysis, each canagliflozin dose (i.e., 100 and 300 mg) will be separately examined.

**Brief Project Background and Statement of Project Significance**

The research project’s background is described in detail in the “Research Proposal” section above. Briefly, the project will address key questions in cardiovascular, metabolic, and renal care by delineating clinical and pathophysiological phenotypes of T2DM in people with established cardiovascular disease.

The proposal will first address whether distinct phenotypes of patients can be identified in patients with T2DM and cardiovascular disease who were enrolled in the CANVAS and CREDENCE trials. Second, it will evaluate whether these clusters have differential outcomes with regards to the trials’ cardiovascular and renal outcomes (e.g., composite of cardiovascular death, nonfatal myocardial infarction, or nonfatal stroke; heart failure hospitalization; cardiovascular death or heart failure hospitalization; cardiovascular death; all-cause mortality; and CKD progression). Third, it will evaluate whether differential treatments effects of canagliflozin exist with each dose of the drug amongst the identified clusters and in the whole cohort. And finally, it will assess the safety of both doses of canagliflozin in each cluster and in the whole cohort, including patients progressing to kidney failure.

The findings of the proposed project will be important as they will allow us to evaluate the feasibility and efficacy of personalized treatment approaches for people with T2DM and established cardiovascular disease. The findings may, in addition, have the potential to guide future clinical trials as examining variations in treatment efficacy and safety across patient clusters may guide the selection of patients who may benefit or respond most to SGLT2 inhibition. Similarly, if some clusters experience increased benefit from SGLT2 inhibition, healthcare systems could leverage our findings to select patients who may derive the most benefit from SGLT inhibition. This may be particularly important in low-income countries or resource-limited settings. Finally, the findings could be utilized to facilitate the design of traditional- and pragmatic
clinical trials. For instance, inclusion criteria may be adapted to include patient clusters at the highest risk of cardiovascular and renal events, enabling parsimonious patient recruitment.

### Specific Aims of the Project

#### Primary Aim:
1. To evaluate if distinct phenotypes of patients with type 2 diabetes can be identified in the CANVAS and CREDENCE trials, through hierarchical latent class analysis.

#### Secondary Aims:
1. To evaluate if these clusters have differential outcomes with regards to CANVAS and CREDENCE’s endpoints (i.e., composite of cardiovascular death, nonfatal myocardial infarction, or nonfatal stroke; heart failure hospitalization; cardiovascular death or heart failure hospitalization; cardiovascular death; all-cause mortality; and CKD progression).

2. To evaluate if these clusters have differential treatment responses to canagliflozin with regards to CANVAS and CREDENCE’s efficacy endpoints (i.e., composite of cardiovascular death, nonfatal myocardial infarction, or nonfatal stroke; heart failure hospitalization; cardiovascular death or heart failure hospitalization; cardiovascular death; all-cause mortality; and CKD progression) and to selected safety outcomes (any adverse event, any serious adverse event, serious adverse event related to canagliflozin, amputation, diabetic ketoacidosis, acute kidney injury, and hyperkalemia). The efficacy and safety of each dose of canagliflozin (i.e., 100 and 300 mg) will be assessed in each cluster and in the whole cohort.

### What is the purpose of the analysis being proposed? Please select all that apply.

- [ ] New research question to examine treatment effectiveness on secondary endpoints and/or within subgroup populations
- [ ] Research on clinical prediction or risk prediction

### Research Methods

The statistical methods are described in detail in the “Statistical Analysis Plan” section below. In brief, latent class analysis defines the distances between subjects based on the combined values of their measured clinical characteristics.10,11

For primary aim #1, we will normalize and reduce the dimensionality of the data by independently transforming categorical and continuous variables with oblique principal
component analysis (PROC VARCLUS procedure in SAS). This will be done to meet the presuppositions of latent class analysis. These data will subsequently be analyzed using polytomous variable latent class analysis with the “poLCA” package in R. The number of classes will be selected based on clinical relevance, the Bayesian information criterion, and the size of the smallest cluster.

For secondary aim #1, the association between cluster membership and clinical outcomes will be assessed utilizing Cox proportional hazards regression models among treatment and placebo assigned patients, respectively. For secondary aim #2, we will utilize interaction terms in the Cox proportional hazards regression model that encompasses the treatment arms to assess whether cluster membership is associated with a differential response to canagliflozin versus placebo, as stratified according to dose (i.e., 100 and 300 mg).

For all Cox proportional hazards regression models, we will adjust for ethnicity, age, sex, New York Heart Association Class, history of hypertension, race, history of myocardial infarction, history of coronary artery disease, history of coronary artery bypass graft surgery, history of peripheral arterial disease, history of heart failure, smoking status, diuretic use, BMI, systolic blood pressure, diastolic blood pressure, estimated glomerular filtration rate, glycosylated hemoglobin, HDL-cholesterol, triglycerides, estimated glomerular filtration rate, urine albumin-to-creatinine ratio, and other available baseline laboratory information (i.e. hemoglobin, electrolytes) and concomitant medications (antihypertensive, antihyperglycemic, antiplatelet, lipid lower, and diuretics), as available.

Software to be used:

- R
- RStudio
- STATA

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study

For primary aim #1, which assesses whether distinct phenotypes exist, all patients enrolled in the CANVAS and CREDENCE trials will be eligible, and clusters will be stratified according to randomization.

For secondary aim #1, which assesses whether clusters have differential outcomes, all patients enrolled in the CANVAS and CREDENCE trials will be eligible, and clusters will be stratified according to randomization.
For secondary aim #2, which assesses whether clusters have differential treatment responses, all patients enrolled in the CANVAS and CREDENCE trials will be eligible, and clusters will be stratified according to randomization and trial drug dose.

Main Outcome Measure and how it will be categorized/defined for your study

The co-primary endpoint of interest for our analysis will be time to heart failure hospitalization, and the composite of cardiovascular death or heart failure hospitalization as defined, respectively, in CANVAS and CREDENCE.

The secondary efficacy endpoints of interest include nonfatal myocardial infarction, or nonfatal stroke; heart failure hospitalization; cardiovascular death; all-cause mortality; regression of albuminuria; two CKD progression-related composite outcomes: the composite outcome of death from renal causes, sustained 40% reduction in eGFR or need for renal replacement therapy; and the composite outcome of doubling serum creatinine, end-stage renal disease, or death from renal causes.

Secondary safety endpoints include any adverse event, any serious adverse event, serious adverse event related to canagliflozin, amputation, diabetic ketoacidosis, acute kidney injury, and hyperkalemia.

Main Predictor/Independent Variable and how it will be categorized/defined for your study

The main predictor/independent variable for all analyses will be the clusters identified with the latent class analysis algorithm. As defined in the “Statistical Analysis Plan” the number of clusters will be selected based on clinical significance (i.e., whether the clusters are clinically distinct in terms of baseline characteristics), the Bayesian information criterion, and the size of the smallest cluster. These criteria has been empirically shown to yield the best results.15

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study

Common variables/risk factors available across the CANVAS and CREDENCE trials will be utilized for the latent class analysis as well any regression models evaluating differential prognosis and treatment response. For the latent class analysis, all baseline variables/risk factors will be put through an oblique principal component analysis (PROC VARCLUS procedure in SAS)12 to normalize and reduce the dimensionality of the data to meet the presuppositions of latent class analysis.13 That is, the SAS procedure will aggregate all available variables into several non-overlapping clusters, which are defined by a summary score characterized by a linear combination of the variables from each patient. (The coefficients of the variable cluster summary score are identified by the first principal component of the variable cluster, which is to be done separately for continuous and categorical variables.) For Cox proportional hazards regression models, the
variables of interest (i.e., baseline clinical, vital, and laboratory characteristics) will be unaltered and defined per the criteria in CANVAS and CREDENCE. Where non-overlapping definitions are observed, clinical experts will be consulted, and the data will be consolidated into logical variables.

Statistical Analysis Plan

Clinical Variable Selection and Data Cleaning

Prior to beginning the analysis, baseline clinical variables will be jointly examined by two study investigators (AS, AR) to assess overlap between CANVAS and CREDENCE. In addition, we will evaluate and remove variables that are highly collinear and/or have limited clinical availability or relevance. We will also remove any variables that are coded as positive in less than 10% of cases as these have been shown to negatively affect patient clustering.\(^{13}\) Following initial variable assessment, multiple imputation (n=5) with the Markov chain Monte Carlo method will be performed if variables have a moderate proportion of missing data (e.g., <40% of data are missing). However, if variables have too high a proportion of missingness (e.g., >40% of data are missing) or if the “missing not at random” assumption is plausible, only complete cases will be used.\(^{16}\)

Data Preparation for Latent Class Analysis

Following clinical variable selection and data cleaning, dimension reduction will be performed to reduce the dimensions of our covariate list. This will be done through independent oblique principal component analysis with the SAS PROC VARCLUS procedure.\(^{12,17}\) In brief, the PROC VARCLUS procedure is an iterative variable clustering process that continuously divides a set of variables into disjoint clusters. That is, “clusters that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters”.\(^{12,17}\) This process will be applied to categorical and continuous variables independently. To determine the appropriate number of variable clusters, we will iteratively evaluate the second eigenvalue of each covariate group. A stopping rule (eigenvalue threshold of 0.7) will be used as suggested by Jackson and colleagues.\(^{18}\) Each patient’s covariate cluster will ultimately be defined by a normalized principal component summary score (i.e., a linear combination of variables) which will be used in the latent class analysis. Analogously, each patient will be defined by a matrix of covariate summary scores allowing for their clustering.

Patient Clustering and Latent Class Analysis

Following data preparation, we will subsequently identify the latent clusters of individuals. This will be done utilizing the “poLCA” package in R, which utilizes a latent class analysis algorithm.\(^{14}\) In brief, latent class analysis is a statistical model that classifies individuals into mutually exclusive (and exhaustive) clusters based on their observed set of measured
characteristics. These clusters will be derived using a maximum likelihood estimation. To avoid finding a local maximum of the log-likelihood function, the model will be estimated 10 times to automate the search for the global maximum. To derive the optimal number of clusters or subgroups, we will evaluate the first minima of the Bayesian information criteria, the size of the smallest class, and the clinical relevance of defined groups. We will use an a priori criteria of at least 200 patients per cluster to promote stability of effect estimates. In addition, we will evaluate model entropy, which is an index of how well the classes are separated. Entropy typically ranges from 0 to 1, and values of 0.8 or greater have been shown to indicate a clinically useful model.

**Descriptive Statistics**

Baseline patient characteristics, labs, vitals, and medications will be described according to the derived patient clusters. Continuous variables will be described as medians and 25th - 75th interquartile ranges. Categorical variables will conversely be described as percentages. Characteristics comparisons across clusters will be done utilizing Kruskal-Wallis test for continuous variables and the Pearson chi-square test for categorical variables.

**Inferential Statistics**

We will first assess the association between cluster membership and clinical outcomes using Cox proportional hazards regression. Proportional hazards assumptions will be assessed by evaluating scaled Schoenfeld residuals. Kaplan-Meier estimated mortality and cause-specific event rates will be plotted according to treatment allocation. Using interaction terms in a Cox regression model, we will also evaluate whether cluster membership is associated with a differential response to randomized therapy. Results will be reported for clinical outcomes as hazard ratios (HR) and 95% confidence intervals (CI) for each cluster in comparison to a reference cluster and an overall p-value is provided to assess the relationship between cluster and outcome. In a sensitivity analysis, we will compare treatment versus placebo effects across each cluster and each dose of canagliflozin in the whole cohort. In models with interactions, an interaction p-value will be provided. We will also provide the estimated HR and 95%CI for canagliflozin therapy versus placebo within cluster subgroups. For all analyses, P<0.05 will be considered significant.

**Project Timeline**

We anticipate completing all study milestones within approximately 8 months, as shown in detail in the Gantt chart (supplemental material). If provided the data, we would start the project on August 15th, 2022. All analyses, including consultations with cardiologists and nephrologists, would be completed in approximately the first 12 weeks. Then, we anticipate spending 6 weeks to write the initial draft of the manuscript and spending another 6 weeks inviting co-authors to help
edit the text. The estimated time of completion of the finalized manuscript would thus be March 15th, 2023, at the latest. We would then develop a journal submission plan, which would include general and specialty medical journals including, but not limited, to the Journal of the American Medical Association (JAMA), the European Heart Journal (EHJ), Circulation, the Clinical Journal of the American Society of Nephrology, and the American Heart Journal. During this time, we would contemporaneously prepare abstracts for major conferences like the American College of Cardiology (ACC) and the Global Cardiovascular Clinical Trialists Forum (CVCT). Once the manuscript is accepted for publication, we would report our results to back to the YODA project by August 15th, 2023, at the latest.

**Dissemination Plan**

Our knowledge translation and dissemination plan is a critical aspect of our proposal as it will ensure that results impact health service delivery, clinical care, and future development of novel therapies. Dr. Abhinav Sharma, principal investigator, currently performs drug evaluations for the Canadian Agency for Drugs And Technologies In Health (CADTH). The results obtained here can influence federal and provincial regulatory/coverage decisions by highlighting specific patient groups that need to be considered when therapies are being evaluated for indications relating to high-risk populations. Dr. Abhinav Sharma is also on the Canadian Cardiovascular Guidelines Committees (Diabetes / Heart Failure), therefore, our results may also impact guideline discussions and recommendations.

As described above, we plan to write manuscript(s) for peer-reviewed medical journals, which we will widely distribute through academic social media and national or international conferences. We would target general and specialty medical journals including, but not limited, to the Journal of the American Medical Association (JAMA), the European Heart Journal (EHJ), Circulation, the Journal and the Clinical Journal of the American Society of Nephrology, the Canadian Journal of Cardiology, and the American Heart Journal. During this time, we would contemporaneously prepare abstracts for major scientific conferences such as the American College of Cardiology (ACC) Annual Scientific Session & Expo, the American Society of Nephrology Kidney Week, and the Global Cardiovascular Clinical Trialists Forum (CVCT).

**Bibliography**


15. Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class


