The YODA Project
Research Proposal Review

The following page contains the final YODA Project review approving this proposal.

The Yale University Open Data Access (YODA) Project
Yale University Center for Outcomes Research and Evaluation (CORE)
195 Church Street, 5th Floor, New Haven, CT 06510

The YODA Project
Research Proposal Review - Final
(Protocol #: 2024-0688)

Reviewers:

- ☐ Nihar Desai
- ☐ Cary Gross
- ☐ Harlan Krumholz
- ☒ Richard Lehman
- ☒ Joseph Ross
- ☒ Joshua Wallach

Review Questions:	**Decision:**
1. Is the scientific purpose of the research proposal clearly described?	Yes
2. Will request create or materially enhance generalizable scientific and/or medical knowledge to inform science and public health?	Yes
3. Can the proposed research be reasonably addressed using the requested data?	Yes, or it's highly likely
4. Recommendation for this data request:	Approve

Comments:

No additional comments

The YODA Project
Research Proposal Review

Revisions were requested during review of this proposal.
The following pages contain the original YODA Project review and the original submitted proposal.

The Yale University Open Data Access (YODA) Project
Yale University Center for Outcomes Research and Evaluation (CORE)
195 Church Street, 5th Floor, New Haven, CT 06510

The YODA Project
Research Proposal Review - Revisions Requested
(Protocol #: 2024-0688)

Reviewers:

- ☐ Nihar Desai
- ☐ Cary Gross
- ☐ Harlan Krumholz
- ☒ Richard Lehman
- ☒ Joseph Ross
- ☒ Joshua Wallach

Review Questions:	**Decision:**
1. Is the scientific purpose of the research proposal clearly described?	Yes
2. Will request create or materially enhance generalizable scientific and/or medical knowledge to inform science and public health?	Yes
3. Can the proposed research be reasonably addressed using the requested data?	Unsure, further clarification from requestor is needed
4. Recommendation for this data request:	Not Approve

Comments:

1. This is an interesting proposal to use two clinical trials to investigate the trans-ethnic effect of canagliflozin on cardiovascular risk in type 2 diabetes using a phenomapping-derived tool to predict individualized risk across East Asian and Caucasian populations. Of note, the authors pre-specify no Main Predictor/Independent Variable or Other Variables of Interest to be used for the analysis. However, do they not intend to use Race/Ethnicity to define the East Asian and Caucasian populations enrolled in the two trials (or are the trials predominantly Caucasian patients, while the Korean EHR data are predominantly East Asian patients)?

2. Similarly, it seems that the phenomapping-derived tool to predict individualized risk uses many different variables. It would be helpful if the authors pre-specified which variables were used for these models.

3. Lastly, depending on the answer to the question above about the Main Predictor/Independent Variable, the use of Korean EHR data is confusing- will those data be uploaded into the YODA Project data sharing platform to be combined for analysis? Or are these data intended to be analyzed using the same methods as the RCT data, but will represent East Asian patients?

Principal Investigator

First Name: Mansu
Last Name: Kim
Degree: Ph.D.
Primary Affiliation: Gwangju Institution of Science and Technology
E-mail: mansu.kim@gist.ac.kr
State or Province: Gwangju-gwangyeoksi
Country: Republic of Korea

General Information

Key Personnel (other than PI):
First Name: Jae-Seung
Last name: Yun
Degree: M.D. Ph.D.
Primary Affiliation: The Catholic University of Korea
SCOPUS ID:
Requires Data Access? Yes

Are external grants or funds being used to support this research?: No external grants or funds are being used to support this research.
How did you learn about the YODA Project?: Scientific Publication

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2024/07/240710-COI-mskim-1.pdf
https://yoda.yale.edu/wp-content/uploads/2024/07/240711-COI-YJS-1.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT01032629 - 28431754DIA3008 - A Randomized, Multicenter, Double-Blind, Parallel, Placebo-Controlled Study of the Effects of JNJ-28431754 on Cardiovascular Outcomes in Adult Subjects With Type 2 Diabetes Mellitus
2. NCT01989754 - 28431754DIA4003 - A Randomized, Multicenter, Double-Blind, Parallel, Placebo-Controlled Study of the Effects of Canagliflozin on Renal Endpoints in Adult Subjects With Type 2 Diabetes Mellitus

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

Trans-ethnic effect of Canagliflozin Cardiovascular risk in Type 2 Diabetes based on phenomapping derived too

Narrative Summary:

We aim to investigate the effect of Canagliflozin on cardiovascular risk across ethnic groups (i.e., East Asian and Caucasian) in Type 2 Diabetes. A previous study developed a predictive tool for drug response using phenomapping from the CANVAS trial. Specifically, the collected key variables, including categorical and continuous, are embedded in a shared latent space to harmonize and calculate distances between individuals. We will then use a phenomapping-derived tool, a powerful tool for predicting individualized risk, to measure individualized hazard ratios within neighborhoods. Finally, we compare individualized risks to examine canagliflozin on cardiovascular risk across ethinic.

Scientific Abstract:

Background: Sodium-glucose cotransporter 2 (SGLT2) inhibitors, such as canagliflozin, have well-documented cardioprotective effects in patients with type 2 diabetes. Despite their benefits, these medications are underused, partially due to high costs and lack of individualized treatment strategies. The differential effects of canagliflozin across diverse ethnic groups, particularly East Asian and Caucasian populations, have not been thoroughly investigated.
Objective: To investigate the trans-ethnic effect of canagliflozin on cardiovascular risk in type 2 diabetes, using a phenomapping-derived tool to predict individualized risk across East Asian and Caucasian populations.
Study design: A retrospective cohort study utilizing dataset of CANVAS trial and external Korean electronic health record (EHR) data to develop and validate a machine learning-based decision support tool. Key variables, both categorical and continuous, will be embedded in a shared latent space to harmonize data and calculate distances between individuals for phenomapping.
Participants: The study includes participants from the CANVAS and CANVAS-R trials, comprising 4,327 and 5,808 patients with type 2 diabetes, respectively. Participants were randomly assigned to receive canagliflozin or placebo. External Korean data were extracted from tertiary hospital-based cohort data and included 5,628 canagliflozin users and 1:1 matched 5,472 non-users.
Primary and secondary outcome: The primary outcome is the time to first major adverse cardiovascular event (MACE), including cardiovascular death, nonfatal myocardial

Brief Project Background and Statement of Project Significance:

Type 2 diabetes is a global health crisis, significantly increasing the risk of cardiovascular diseases (CVD), which are leading causes of morbidity and mortality. Sodium-glucose cotransporter 2 (SGLT2) inhibitors, such as canagliflozin, have demonstrated substantial cardioprotective effects. Despite their proven benefits, these medications are underutilized due to high costs and the lack of tailored treatment strategies that address the diverse responses across different ethnic groups.
In this research, we aim to address this gap by investigating the trans-ethnic effects of canagliflozin on cardiovascular risk in type 2 diabetes among East Asian and Caucasian populations. Using a novel phenomapping-derived tool, we will embed categorical and continuous key variables for cardiovascular risk factors into a shared latent space to calculate the distance between individuals and predict individualized risk. This approach will help in understanding the differential impact of canagliflozin across ethnic groups, ultimately leading to more personalized and effective treatment strategies.
The significance of this project lies in its potential to enhance generalizable scientific and medical knowledge by providing insights into how canagliflozin affects different ethnic populations. By identifying specific patient phenotypes that benefit most from this medication, we can optimize its use, improve clinical outcomes, and inform public health strategies. This work will contribute to the broader goal of precision medicine, ensuring that therapeutic decisions are better tailored to

individual patient profiles, thus maximizing the benefits of treatment and resource utilization.

Specific Aims of the Project:

To investigate the trans-ethnic effect of canagliflozin on cardiovascular risk in type 2 diabetes, using a phenomapping-derived tool to predict individualized risk across East Asian and Caucasian populations.

Study Design:

Individual trial analysis

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

Confirm or validate previously conducted research on treatment effectiveness

Research on clinical prediction or risk prediction

Research Methods

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

The data for this study will be sourced from the CANVAS (Canagliflozin Cardiovascular Assessment Study) and CANVAS-R (CANVAS-Renal) trials, obtained through the Yale University Open Data Access (YODA) Project. Additionaly we will correct data from Catholic University of Korea. These trials include detailed baseline variables and follow-up data for patients with type 2 diabetes. Inclusion Criteria for the CANVAS Trial are 1) Age: Patients aged 30 years or older with established atherosclerotic cardiovascular disease (ASCVD) or patients aged 50 years or older with two or more ASCVD risk factors, 2) Diagnosis: Patients with a diagnosis of type 2 diabetes, with HbA1c levels between 7.0% and 10.5% (53–91 mmol/mol). 3) ASCVD Risk Factors: At least two of the following risk factors. and Exclusion Criteria are subject who has Type 1 diabetes, Severe illness, History of urinary tract infection (UTI) or genital tract infection (GTI), Diabetic ketoacidosis (DKA), Hypersensitivity to canagliflozin or its components, Pregnancy and lactation, and Life expectancy of less than two years.

Primary and Secondary Outcome Measure(s) and how they will be categorized/defined for your study:

The primary outcome measure for this study is the time to the first occurrence of a composite of major adverse cardiovascular events. This composite includes Cardiovascular death, nonfatal myocardial infarction, and Nonfatal stroke. Secondary outcomes include hospitalization for heart failure and progression of renal disease.

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

Variables collected by both institutions are considered in this study. Specifically, variables with >20% missingness and >0.7 collinearity of variables are removed. We imputed missing data using chained random forests with predictive mean matching).

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

None

Statistical Analysis Plan:

First, we plan to propose a modified Gower's method to find the phenotypic similarity of all individuals. Specifically, we will use the weighted Gower method, which computes a metric of dissimilarity between two data points, including both numerical (e.g., age, BMI) and non-numerical data (e.g., history of hypertension, smoking status). The weighting in the Gower method will be adjusted to give appropriate importance to different types of variables based on their clinical relevance and statistical properties.

Second, once the phenotypic similarities are calculated, individuals will be grouped into phenotypically similar clusters. These clusters will be defined as neighborhoods encompassing 5% to 30% of the most similar individuals to each index individual. Within these neighborhoods, we will conduct multivariate Cox regression models to estimate individualized hazard ratios (HR) for major adverse cardiovascular events (MACE). The resulting HRs will be mapped onto a high-dimensional space and visualized using t-SNE (t-distributed Stochastic Neighbor Embedding). This visualization will allow us to compare the HR maps between East Asian and Caucasian populations to identify any significant differences in risk patterns.

Third, we will develop an individualized prediction model using a penalized regression approach, such as LASSO (Least Absolute Shrinkage and Selection Operator) or elastic net regression. This model will be trained to predict cardiovascular risk based on the phenotypic data. The model's performance will be rigorously evaluated using nested five-fold cross-validation to ensure robustness and generalizability. To investigate the trans-ethnic effect of canagliflozin, we will apply the individualized risk prediction model developed from the Caucasian population to the East Asian population and vice versa. This cross-application will help us understand how well the risk factors and treatment effects generalize across different ethnic groups.

For model interpretability, we will use SHAP (Shapley Additive Explanations) values to assess feature importance. SHAP values provide a unified measure of the contribution of each feature to the prediction, making it easier to interpret the influence of individual variables on the predicted cardiovascular risk. This will help us identify which phenotypic characteristics are most critical in determining the individualized risk and how they differ between ethnic groups.

By following this detailed and rigorous approach, we aim to uncover important insights into the trans-ethnic effects of canagliflozin on cardiovascular risk in type 2 diabetes, ultimately contributing to more personalized and effective treatment strategies.

Software Used:

Python

Project Timeline:

In our study, data cleaning and preprocessing will be conducted over three months. Based on the phenomapping-derived tool, the model will be developed, trained, and tested over the following four months. Interpretation of results and sensitivity analyses will be conducted, and manuscript drafting will occur in the subsequent two months, with the first submission planned for nine months after the project begins. Extensions will be requested if additional time is needed.

Dissemination Plan:

We plan to prepare and submit multiple manuscripts detailing the study's findings. The primary manuscript will focus on the overall trans-ethnic effects of canagliflozin on cardiovascular risk in type 2 diabetes. Additional manuscripts may explore specific aspects of the data, such as subgroup analyses and detailed methodological approaches. Potentially Suitable Journals are Diabetes Care, The Lancet Diabetes & Endocrinology, Journal of the American Medical Association (JAMA), The New England Journal of Medicine (NEJM), Circulation, and European Heart Journal.

Bibliography:

Oikonomou EK, Suchard MA, McGuire DK, Khera R. Phenomapping-Derived Tool to Individualize the Effect of Canagliflozin on Cardiovascular Risk in Type 2 Diabetes. Diabetes Care 2022;45:965-74.

Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondu N, Shaw W, Law G, Desai M, Matthews DR. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. N Engl J Med 2017;377:644-57.

Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971;27:857–871

McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. Accessed 28 January 2022. Available from https://arxiv.org/abs/1802.03426

Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Accessed 28 January 2022. Available from https://arxiv.org/abs/1603.02754 25. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67