

Principal Investigator

First Name: Joshua

Last Name: Wallach

Degree: PhD, MS

Primary Affiliation: Rollins School of Public Health, Emory University

E-mail: joshua.wallach@emory.edu

State or Province: GA

Country: United States

General Information

Key Personnel (other than PI):

First Name: Joseph

Last name: Ross

Degree: MD, MHS

Primary Affiliation: Yale School of Medicine

SCOPUS ID:

Requires Data Access? Yes

First Name: Luis

Last name: Correia

Degree: MD

Primary Affiliation: Rollins School of Public Health, Emory University

SCOPUS ID:

Requires Data Access? Yes

First Name: Erfan

Last name: Taherifard

Degree: MD, MPH

Primary Affiliation: Yale School of Medicine

SCOPUS ID:

Requires Data Access? Yes

First Name: Maya

Last name: Deshmukh

Degree: MPH

Primary Affiliation: Rollins School of Public Health, Emory University

SCOPUS ID:

Requires Data Access? Yes

Are external grants or funds being used to support this research?: External grants or funds are being used to support this research.

Project Funding Source: Johnson & Johnson

How did you learn about the YODA Project?: Colleague

Conflict of Interest

<https://yoda.yale.edu/wp-content/uploads/2025/05/JRoss-COI.pdf>

<https://yoda.yale.edu/wp-content/uploads/2025/05/JWallach-COI.pdf>

<https://yoda.yale.edu/wp-content/uploads/2025/05/Yoda-COI-Form-Luis.pdf>

<https://yoda.yale.edu/wp-content/uploads/2025/05/COI-FORM-ET.pdf>

<https://yoda.yale.edu/wp-content/uploads/2025/05/COI-FORM-MD.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. [NCT02139943 - 28431754DIA2004 - A Randomized Phase 2, Double-blind, Placebo-controlled, Treat-to-Target, Parallel-group, 3-arm, Multicenter Study to Assess the Efficacy and Safety of Canagliflozin as Add-on Therapy to Insulin in the Treatment of Subjects With Type 1 Diabetes Mellitus](#)
2. [NCT01106651 - 28431754DIA3010 - A Randomized, Double-Blind, Placebo-Controlled, Parallel-Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin Compared With Placebo in the Treatment of Older Subjects With Type 2 Diabetes Mellitus Inadequately Controlled on Glucose Lowering Therapy](#)
3. [NCT00642278 - 28431754DIA2001 - A Randomized, Double-Blind, Placebo-Controlled, Double-Dummy, Parallel Group, Multicenter, Dose-Ranging Study in Subjects With Type 2 Diabetes Mellitus to Evaluate the Efficacy, Safety, and Tolerability of Orally Administered SGLT2 Inhibitor JNJ-28431754 With Sitagliptin as a Reference Arm](#)
4. [NCT00968812 - 28431754DIA3009 - A Randomized, Double-Blind, 3-Arm Parallel-Group, 2-Year \(104-Week\), Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of JNJ-28431754 Compared With Glimepiride in the Treatment of Subjects With Type 2 Diabetes Mellitus Not Optimally Controlled on Metformin Monotherapy](#)
5. [NCT01064414 - 28431754DIA3004 - A Randomized, Double-Blind, Placebo-Controlled, 3-Arm, Parallel-Group, 26-Week, Multicenter Study With a 26-Week Extension, to Evaluate the Efficacy, Safety and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus Who Have Moderate Renal Impairment](#)
6. [NCT01081834 - 28431754DIA3005 - A Randomized, Double-Blind, Placebo-Controlled, Parallel-Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin as Monotherapy in the Treatment of Subjects With Type 2 Diabetes Mellitus Inadequately Controlled With Diet and Exercise](#)
7. [NCT01106625 - 28431754DIA3002 - A Randomized, Double-Blind, Placebo-Controlled, 3-Arm, Parallel-Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin and Pioglitazone Therapy](#)
8. [NCT01106677 - 28431754DIA3006 - A Randomized, Double-Blind, Placebo and Active-Controlled, 4-Arm, Parallel Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin Monotherapy](#)
9. [NCT01106690 - 28431754DIA3012 - A Randomized, Double-Blind, Placebo-Controlled, 3-Arm, Parallel-Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin and Sulphonylurea Therapy](#)
10. [NCT01137812 - 28431754DIA3015 - A Randomized, Double-Blind, Active-Controlled, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin Versus Sitagliptin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin and Sulphonylurea Therapy](#)
11. [NCT01340664 - 28431754DIA2003 - A Randomized, Double-Blind, Placebo-Controlled, 3-Arm, Parallel-Group, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin](#)
12. [NCT01381900 - 28431754DIA3014 - A Randomized, Double-Blind, Placebo-Controlled, Parallel Group, 18-Week Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on](#)

- [Metformin Alone or in Combination With a Sulphonylurea](#)
13. [NCT01809327 - 28431754DIA3011 - A Randomized, Double-Blind, 5-Arm, Parallel-Group, 26-Week, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in Combination With Metformin as Initial Combination Therapy in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control With Diet and Exercise](#)
 14. [NCT02025907 - 28431754DIA4004 - A Randomized, Double-blind, Placebo Controlled, 2-arm, Parallel-group, 26-week, Multicenter Study to Evaluate the Efficacy, Safety, and Tolerability of Canagliflozin in the Treatment of Subjects With Type 2 Diabetes Mellitus With Inadequate Glycemic Control on Metformin and Sitagliptin Therapy](#)
 15. [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](#)
 16. [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](#)
 17. [NCT00650806 - 28431754OBE2001 - A Randomized, Double-Blind, Placebo-Controlled, Parallel-Group, Dose-Ranging Study to Investigate the Safety and Efficacy of JNJ-28431754 in Nondiabetic Overweight and Obese Subjects](#)
 18. [NCT02243202 - 28431754OBE2002 - A Randomized, Double-Blind, Placebo-Controlled, Parallel-Group Study to Investigate the Safety and Efficacy of the Co-administration of Canagliflozin 300 mg and Phentermine 15 mg Compared With Placebo for the Treatment of Non-diabetic Overweight and Obese Subjects](#)
 19. [NCT02065791 - 28431754DNE3001 - A Randomized, Double-blind, Event-driven, Placebo-controlled, Multicenter Study of the Effects of Canagliflozin on Renal and Cardiovascular Outcomes in Subjects With Type 2 Diabetes Mellitus and Diabetic Nephropathy](#)

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

Research Proposal

Project Title

Agreement of Treatment Effects in a Meta-Analysis Using Real and Synthetic Individual Participant-Level Data

Narrative Summary:

This study will examine how treatment effects in meta-analyses may differ when using real versus synthetic individual participant-level data (IPD), focusing on clinical trials evaluating the efficacy and safety of canagliflozin for treating type 2 diabetes mellitus. Primary efficacy outcome measures will include HbA1c levels and major cardiovascular events, while safety outcome measures will include serious adverse events. Major adverse cardiovascular events will be defined as a combined endpoint of cardiovascular death, non-fatal infarction, and non-fatal stroke.

Scientific Abstract:

Background: Synthetic IPD have been proposed as a potential solution to mitigate the risk of re-identification in data sharing. However, further research is needed to evaluate the impact of using synthetic data from multiple trials in meta-analyses.

Objective: To examine how treatment effects in meta-analyses may differ when using real versus synthetic IPD, focusing on clinical trials evaluating the efficacy and safety of canagliflozin for treating type 2 diabetes mellitus.

Study design: Meta-analyses and case study

Participants: Participants aged ≥ 18 years with a diagnosis of type 2 diabetes mellitus without distinction in terms of baseline body mass index or baseline HbA1c concentrations.

Main outcome measure: Co-primary efficacy: HbA1c levels and major cardiovascular events; Safety: serious adverse events.

Statistical analysis: We will use two data sources, real and synthetic IPD, to perform a series of two-stage meta-analyses. In the first stage, we will calculate trial specific mean differences and hazard ratios and their corresponding 95% confidence intervals. In the second stage, effect estimates from each individual trial will be combined using a random effects model estimated via restricted maximum likelihood estimation. We will derive confidence intervals using the Hartung Knapp approach. We will also conduct a one-stage meta-analysis to produce effect estimates and 95% confidence intervals. We will compare the individual and summary treatment effects from the real and synthetic IPD meta-analyses based on their significance ($P < 0.05$ vs. $P \geq 0.05$) and direction.

Brief Project Background and Statement of Project Significance:

Efforts to expand access to clinical trial data have grown, focusing on both summary data (e.g., protocols, clinical study reports, and publications) and individual participant-level data (IPD). Many stakeholders have recognized the benefits of sharing IPD [1-6], such as enhancing the reliability of scientific findings, reducing redundant experiments, maximizing the contributions of participants, and enabling secondary analyses. De-identified IPD from thousands of clinical trials are now available for request through various data-sharing platforms (e.g., Biological Specimen and Data Repository Information Coordinating Center [BioLINCC], ClinicalStudyDataRequest.com, Project Data Sphere, and the Yale Open Data Access [YODA] Project). These platforms have enabled hundreds of secondary analyses, including validation, subgroup, replication, and meta-analysis studies [7]. However, concerns persist regarding the sharing of IPD [8-10]. Although data anonymization can be employed to protect privacy and ensure compliance with ethical and legal standards, patient confidentiality can never be fully guaranteed [8,9]. While higher levels of data de-identification may reduce the risk of re-identification, these measures can be resource-intensive and may compromise the utility of the data, highlighting the need for alternative approaches [10].

To mitigate the risk of re-identification in data sharing, synthetic data have been proposed as a potential solution. Machine learning models can be trained on real IPD to capture key characteristics and generate new synthetic data that retain the statistical properties of the original dataset, including sample size, participant rows, and variable columns. Since synthetic data do not have a direct one-to-one mapping to the original IPD, they are considered to pose lower privacy risks [11]. The concepts and methods used to generate synthetic data have existed for decades, but these techniques have not been widely evaluated and adopted in medical research.

Previous studies have evaluated the feasibility of generating synthetic versions of clinical trial data, compared the results and conclusions from analyses using real versus synthetic IPD, and examined the ability to re-identify participants. These evaluations suggest that several promising approaches exist to generate synthetic data that closely resemble the original datasets. In particular, two oncology case studies found that analysis results and conclusions were similar when repeated using both real and synthetic data, with low re-identification risk [11,12]. However, these efforts have primarily focused on evaluating the performance of synthetic data in individual trial results. Further research is needed to assess the impact of using synthetic data from multiple trials in meta-analyses, including evaluating whether summary efficacy and safety outcome estimates are consistent when derived from real versus synthetic IPD.

Specific Aims of the Project:

Specific Aim 1: To conduct meta-analyses comparing canagliflozin to placebo, focusing on primary efficacy outcomes (HbA1c reduction and major cardiovascular events), and to evaluate the concordance between summary effect estimates generated using real versus synthetic IPD.

Specific Aim 2: To conduct meta-analyses comparing canagliflozin to placebo, focusing on safety outcomes (serious adverse events), and to evaluate the concordance between summary effect estimates generated using real versus synthetic IPD.

Study Design:

Meta-analysis (analysis of multiple trials together)

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

Participant-level data meta-analysis

Meta-analysis using only data from the YODA Project

Develop or refine statistical methods

Other: Research of meta-analysis methods

Research Methods

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

Real IPD: We will identify and request access to the IPD from RCTs comparing canagliflozin (50-300 mg), with or without another active treatment, versus a placebo from the YODA Project platform. Eligible RCTs will include people (aged ≥ 18 years) with a diagnosis of type 2 diabetes mellitus without distinction in terms of baseline body mass index or baseline HbA1c concentrations. RCTs need to have assessed at least one outcome among the following: HbA1c, major adverse cardiovascular events, and serious adverse events in the double-blind study period.

Synthetic IPD: For the requested RCTs, Johnson & Johnson will work with a data synthesization vendor to generate synthetic IPD for each RCT. The method that will be used will sequentially apply classification and regression trees (CART) to generate synthetic data by modeling the relationships between variables [1].

1. Emam KE, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. *J Am Med Inform Assoc.* Jan 15 2021;28(1):3-13.

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

Co-Primary: HbA1c difference from baseline and weeks 12, 18, 26, and 52; time to occurrence of the first major adverse cardiovascular event. For HbA1c data, if an observation is missing at a time point (+/-3 weeks; except for baseline, where only measures at --3 weeks will be considered), it will be replaced using the last observation carried forward method.

Major adverse cardiovascular events will be defined as a combined endpoint of cardiovascular death, non-fatal infarction, and non-fatal stroke.

Secondary: Time to occurrence of the first serious adverse event. We will follow the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use's definition of serious adverse event.

Effect estimates will be expressed in terms of mean differences for changes in HbA1c levels and hazard ratios for major adverse cardiovascular events and serious adverse events.

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

Canagliflozin (50-300 mg), with or without another active treatment, versus a placebo.

HbA1c: Mean difference for HbA1c; MACE: Hazard ratio; Serious adverse event: Hazard ratio.

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

To characterize the sample:

Age (median, IQR)

% Female

Statistical Analysis Plan:

Synthetic IPD: The method that will be used will sequentially apply classification and regression trees (CART) to generate synthetic data by modeling the relationships between variables V1, V2, and V3 [1]. Decision trees will be constructed using a greedy algorithm that recursively splits the variables, selecting the one that minimizes a loss function at each step. The goal is to find optimal binary splits for each variable. Once the trees are built, pruning will be applied using a cost-complexity criterion to prevent overfitting. The synthesis process begins by sampling from the actual or fitted distribution of the first variable, V1, to create a synthetic version, sV1. This synthetic value, sV1, is then input into the first tree model, which predicts a distribution for the second variable, V2. The synthetic value for V2, sV2, is either sampled from this predicted distribution or smoothed using a kernel density estimator to introduce variability. Next, sV1 and sV2 are input into the second tree model to predict the distribution for V3, generating the synthetic value sV3. This process will continue for all variables in the dataset, ensuring the synthetic data reflect the original relationships between variables.

Meta-analyses: For each endpoint and analysis, we will consider two data sources: Real and synthetic IPD. We will perform a series of two stage meta-analyses. In the first stage, we will calculate trial specific mean differences and hazard ratios and their corresponding 95% confidence intervals. In the second stage, effect estimates from each individual trial will be combined using a random effects model estimated via restricted maximum likelihood estimation. We will derive confidence intervals using the Hartung Knapp approach. We will also conduct a one-stage meta-analysis (e.g., a one-stage generalized fixed and random study-specific model using the Simmons and Higgins method with random study-specific effects, implemented with the lme4 package in R) to produce effect estimates and 95% confidence intervals.

Comparisons between real and synthetic IPD meta-analyses: Individual (study specific) and summary treatment effects will be characterized on the basis of their significance (that is, $P < 0.05$ vs. $P \geq 0.05$) and direction (that is, increased for hazard ratios greater than 1 or mean differences greater than 0, and decreased for hazard ratios and mean differences less than 0). Treatment effect estimates for the same endpoints will be classified as concordant if the direction of the treatment effect estimates is concordant and both the treatment effect estimates are significant, or if the effect estimates are both not significant. Treatment effect estimates that do not fulfill either of these criteria will be classified as discordant. Although P values are imperfect measures, our binary classification system, based on the traditional alpha cut-off value of 0.05, is useful for showing how significance is most often defined in the literature.

We will consider $P < 0.05$ to be significant for all two-sided tests. All analyses will be done using the meta package in R (version 4.1.2).

1. Emam KE, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. *J Am Med Inform Assoc.* Jan 15 2021;28(1):3-13.

Software Used:

R

Project Timeline:

Data request: May 2025
Analysis start: September 2025
Project finish: July 2026
Complete manuscript: September 2026

Dissemination Plan:

We will publish our evaluation in a peer-reviewed medical or methodological journal (E.g. BMC Medical Research Methodology).

Bibliography:

1. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) Project—A Mechanism for Data Sharing. *N Engl J Med*. Aug 2016;375(5):403-5. doi:10.1056/NEJMp1607342
2. Kaiser J. NIH aims to beef up clinical trial design as part of new data sharing rules. *Science* 2016.
3. Pencina MJ, Louzao DM, McCourt BJ, et al. Supporting open access to clinical trial data for researchers: The Duke Clinical Research Institute-Bristol-Myers Squibb Supporting Open Access to Researchers Initiative. *Am Heart J*. Feb 2016;172:64-9. doi:10.1016/j.ahj.2015.11.002
4. Taichman DB, Backus J, Baethge C, et al. Sharing Clinical Trial Data—A Proposal from the International Committee of Medical Journal Editors. *N Engl J Med*. Jan 2016;374(4):384-6. doi:10.1056/NEJMe1515172
5. Institute of Medicine (IOM). *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risks*. Washington,DC: The National Academies Press. 2015.
6. National Academies of Sciences, Engineering, and Medicine. 2020. *Reflections on Sharing Clinical Trial Data: Challenges and a Way Forward: Proceedings of a Workshop*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/25838>.
7. Vazquez E, Gouraud H, Naudet F, et al. Characteristics of available studies and dissemination of research using major clinical data sharing platforms. *Clin Trials*. 12 2021;18(6):657-666. doi:10.1177/17407745211038524
8. Kokosi T, Harron K. Synthetic data in medical research. *BMJ Med*. 2022;1(1):e000167. doi:10.1136/bmjmed-2022-000167
9. Abgrall G, Monnet X, Arora A. Synthetic Data and Health Privacy. *JAMA*. Feb 18 2025;333(7):567-568. doi:10.1001/jama.2024.25821
10. Tucker K, Branson J, Dilleen M, et al. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med Res Methodol*. Jul 08 2016;16 Suppl 1(Suppl 1):77. doi:10.1186/s12874-016-0169-4
11. El Kababji S, Mitsakakis N, Fang X, et al. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. *JCO Clin Cancer Inform*. Sep 2023;7:e2300116. doi:10.1200/CCI.23.00116
12. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, Collaborators G-F. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open*. Apr 16 2021;11(4):e043497. doi:10.1136/bmjopen-2020-043497