Drugs - Real World Outcomes (2023) 10:235–247
https://doi.org/10.1007/s40801-023-00353-7

ORIGINAL RESEARCH ARTICLE

An Attempt to Replicate Randomized Trials of Diabetes Treatments Using a Japanese Administrative Claims and Health Checkup Database: A Feasibility Study

Ryozo Wakabayashi[1,2] · Takahiro Hirano[1,2] · Tadashi Koga[1,2] · Ryohei Kuwatsuru[2,3]

Accepted: 8 January 2023 / Published online: 1 February 2023

Abstract

Background Use of real-world evidence (RWE) has been limited for evaluating effectiveness because of the lack of confidence in its reliability. Examining whether a rigorously designed observational study using real-world data (RWD) can reproduce the results of a randomized controlled trial (RCT) will provide insights into the implementation of high-quality RWE studies that can produce valid conclusions.

Objective We aimed to replicate published RCTs using a Japanese claims and health checkup database and examine whether the emulated RWE studies' results agree with those of the original RCTs.

Methods We selected three RCTs on diabetes medications for replication in patients with type 2 diabetes. The study outcome was either the change or percentage change in HbA1c levels from baseline. We designed three observational studies using the RWD to mimic the critical study elements of the respective RCTs as closely as possible. We performed 1:1 propensity score nearest-neighbor matching to balance the groups for potential confounders. The differences in outcomes between the groups and their 95% confidence intervals (CIs) were calculated in each RWE study, and the results were compared with those of the RCT.

Results Patient characteristics, such as age, sex, and duration of diabetes, differed between the RWE studies and RCTs. In Trial 1 emulation, the percentage changes in HbA1c levels were larger in the treatment group than in the comparator group (difference −6.21, 95% confidence interval (CI) −11.01 to −1.40). In Trial 2, the change in HbA1c level was larger in the treatment group (difference −0.01; 95% CI −0.25 to 0.23), and in Trial 3, it was smaller in the treatment group (difference 0.46; 95% CI −0.01 to 0.94). These results did not show regulatory or estimate agreement with the RCTs.

Conclusions None of the three emulated RWE studies using this claims and health checkup database reproduced the same conclusions as the RCTs. These discrepancies could largely be attributed to design differences between RWE studies and RCTs, primarily due to the lack of necessary data in the database. This particular RWD source may not be the best fit for evaluating treatment effects using laboratory data as the study outcome.

1 Introduction

Real-world evidence (RWE) is "the clinical evidence about the usage and potential benefits and risks of a medical product derived from analysis of real-world data (RWD)," the routinely collected health-related data [1]. Although randomized controlled trials (RCTs) are considered the "gold standard" for evaluating treatment effects and safety [2], their highly selective populations and tightly controlled settings limit their generalizability. RWE can supplement the evidence obtained from RCTs by providing information on their effectiveness in clinical settings [3]. In this sense, RCTs and RWE should be regarded as mutually complementary rather than competing relationships [4]. Furthermore, RWE can be utilized throughout the life cycle of a drug, let alone for effectiveness and safety evaluation [5, 6], which is expected to accelerate the drug development process.

Despite its great potential, the use of RWE remains limited, especially at the contribution level to regulatory decision-making [7, 8]. RWE studies lack randomization and primary data collection, and people are concerned about their drawbacks, such as low data quality and improper analytical methods [9, 10]. These factors complicate the interpretation of causal inference in these studies, leading to less confidence in the reliability of RWE [2]. Given this situation, enhancing people's trust in RWE is crucial for facilitating its use, especially in

Extended author information available on the last page of the article

Key Points

We attempted to replicate the published randomized controlled trials (RCTs) on diabetic medications in patients with type 2 diabetes in Japan using a Japanese claims and health checkup database.

We closely designed three observational studies using this real-world data (RWD) source, mimicking the critical study elements of RCTs; however, various design elements could not be precisely emulated, primarily because of the lack of necessary data.

This particular RWD source may not be the best fit for these specific research questions, requiring laboratory data as study outcomes. More RCT replication exercises should be conducted to accumulate knowledge on the opportunities and limitations of real-world evidence studies.

effectiveness evaluation. For this purpose, first, when we can obtain valid conclusions from RWE studies instead of RCTs, and second, how it can be implemented should be identified [2].

To obtain insights into those "when" and "how," efforts have been made to replicate RCTs results with rigorously designed observational studies using RWD [1]. These are attempts to replicate an RCT by mimicking its critical study elements (e.g., study population, treatments, outcomes) and comparing the results between the RCT and emulated RWE study. Such replication exercises may provide insights into clinical scenarios (e.g., indications and outcomes), study designs, and analytical approaches for implementing high-quality RWE studies that can produce valid conclusions [7]. The RCT DUPLICATE Initiative—a collaboration project of Brigham and Women's Hospital, Harvard Medical School, the U.S. Food and Drug Administration (FDA), and Aetion—is one such leading project aimed at replicating 30 completed phase III or IV RCTs using health claims data [11, 12]. Some other attempts also existed, which were not only to replicate completed RCTs [13] but also to predict the results of ongoing RCTs [14, 15], albeit mainly in the USA.

However, such attempts to replicate RCTs have not yet been made in Japan. RWD is increasingly being used in Japan for drug safety assessment and epidemiological research. Still, RWE studies have yet to be entirely acknowledged to contribute to decision-making on effectiveness, as in other countries, due to concerns about its reliability [5]. Thus, more knowledge on their opportunities and limitations should be accumulated through RCT replication exercises using Japanese RWD. This will enhance people's confidence in RWE and facilitate its use in Japan. Despite previous overseas practices, such country-specific attempts are essential because healthcare systems/policies and available RWD sources vary among countries.

Therefore, in this study, we attempted to replicate published RCTs using the JMDC database, one of Japan's most commonly used commercial databases [16]. We chose diabetes studies for replication because the increasing disease burden of diabetes is a serious public health concern in Japan [17], where one in eight adults has diabetes [18]. The proper management of diabetes is an important clinical mission. The JMDC database contains claims and health checkup data, including blood test results [19]. The availability of health checkup data is a unique characteristic of this database, which enabled us to target diabetes studies with hemoglobin A1c (HbA1c) as the study outcome, which cannot be emulated using RWD sources such as administrative databases.

2 Methods

2.1 Study Overview

This was a feasibility study to examine whether RWE studies using Japanese RWD can reproduce the results of published, specific RCTs, if closely designed. After selecting RCTs of a particular clinical area for replication, we designed RWE studies to mimic the trials' critical study elements, such as inclusion/exclusion criteria, interventions, comparators, outcomes, covariates, and follow-up periods, as precisely as possible, and analyzed treatment effectiveness. We then compared the obtained results between the RCTs and the emulated RWE studies. This study was approved by the ethics committee of Juntendo University, Tokyo, Japan (E21-0284).

2.2 Selection of Randomized Controlled Trials (RCTs) for Replication

We targeted RCTs that evaluated the efficacy of diabetic medications on HbA1c levels (not necessarily as primary outcomes) in patients with diabetes in Japan, published in the last 10 years, and potentially replicable with our RWD source. Figure S1 of the Electronic Supplementary Material (ESM) shows the flow chart of the RCT selection process. We searched PubMed on 1 June, 2022, using the following search terms: ("diabetes"[Title/Abstract] AND "Japanese"[Title/Abstract] AND "HbA1c"[Title/Abstract])

AND ((y_10[Filter]) AND (randomized controlled trial [Filter])).

Of the 149 articles obtained, those unsuitable for our target (e.g., non-RCTs, no HbA1c outcomes, and non-Japanese patients included) were excluded after reviewing the titles and abstracts. We also excluded placebo-controlled trials and studies with complicated designs or treatment schemes, making them non-replicable with RWD. Thus, the studies were limited to active-controlled RCTs with simple treatment schemes. Additionally, studies that did not find statistically significant differences in HbA1c outcomes were excluded. This was because, in the case of null results, the agreement between RWE studies and RCTs is more likely to occur because of measurement error, given that misclassification in RWE studies can result in a bias toward the null [11]. The full exclusion criteria are presented in Fig. S1 of the ESM.

This screening process resulted in 13 candidate RCTs, for which we examined their feasibility for replication with RWD. Within the database, we identified patients who had (1) necessary prescription records (study drugs or comparator drugs) and (2) HbA1c data within 90 days before the first prescription and 180 days after the trial follow-up period. If the number of patients who met these two minimum conditions was already less than 100 in either of the groups, the ultimate number of patients meeting all the study's inclusion/exclusion criteria would be minimal. Therefore, these RCTs were considered unsuitable to replicate using this RWD source; thus, they were excluded from candidates.

Consequently, we obtained three RCTs, all in type 2 diabetes, for replication: Trial 1 compared ipragliflozin (sodium-dependent glucose transporter-2 inhibitor [SGLT-2i]) versus metformin (biguanides) [20]; Trial 2 compared sitagliptin (dipeptidyl peptidase-4 inhibitors (DPP-4i)) and pioglitazone (thiazolidinediones) [21]; and Trial 3 compared insulin degludec/insulin aspart versus insulin glargine [22]. Summaries of these RCTs (Trials 1–3) are presented in Table 1.

2.3 Data Source

This study used the JMDC database, which consists of claims and health checkup results of insured employees and their dependents, collected from health insurance societies [19]. In Japan, people usually undergo a health checkup annually because employers must provide employees with a yearly health checkup under the Industrial Safety and Health Act, and the insurers are obligated to provide an annual health checkup ("specific health checkup" aiming to prevent metabolic syndrome) to their insurers and their dependents aged ≥ 40 years.

The database includes the following information: patient attributes (age and sex), diagnoses, medical care activities, prescriptions (date, dose, and supply days), and health checkup results (including body mass index (BMI), blood measures, and lifestyle habits). Data have been anonymized, but personal IDs enable tracking the same individuals across different hospitals, as long as the same insurance society covers them. This traceability is one advantage of this database for use in research on chronic diseases, such as diabetes. Indeed, many RWE studies have been conducted in diabetes research using this database [23].

The present study used data of patients with type 2 diabetes (10th revised version of the International Statistical Classification of Diseases (ICD-10) codes, E11-14) between January 2005 and April 2020.

2.4 Replication of Three Real-World Data (RCTs) Using RWD

We designed three observational studies using RWD (RWE studies), mirroring the critical study elements of the respective RCTs, to emulate these target RCTs.

2.4.1 Population

In the emulation of each RCT, data for patients with prescriptions for study treatment (study drugs or comparator drugs) were extracted from the database. The cohort entry date (CED) was defined as the first prescription date of the study treatment, that is, treatment initiation. We conditioned patients to have data for at least 180 days before CED to check for previous treatment status and to extract new users of the study treatment. The patients also had to have the necessary data during the baseline and post-treatment assessment windows. Therefore, patients without health-checkup data within 90 days before CED and within 180–360 days after CED were excluded.

The other inclusion/exclusion criteria, which were defined to mirror those of the corresponding RCT as closely as possible, were applied to these patients, unless the criterion was not imitable with our RWD. The original patient criteria in the RCTs and the corresponding operational definitions in our emulations are provided in Table S1 of the ESM. However, in two of our emulation studies, applying all imitable patient criteria resulted in almost no patients (0 or 5). In this case, the patient criterion that most affected the number of patients, that is, the criterion regarding antidiabetic medications before CED, was disregarded to secure the number of patients. The modified definitions are presented in Table S1 of the ESM and the number of subjects eliminated based on each criteria were presented in Fig. S2 of the ESM.

Table 2 illustrates the specification and emulation of a key component of the target trial. Overall, the timing when

Table 1 Summary of three randomized controlled trials (RCTs) for replication

Study		Design	Study population	Follow-up	Group		Outcome measure	Mean difference in outcomes [95% CI] (treatment −comparator)
No.	Author, year				Treatment	Comparator		
Trial 1	Koshizaka et al, 2019 [20]	2-arm RCT	Japanese adults with T2D treated with sitagliptin	24 wk	Ipragliflozin 50 mg/day (n = 48)	Metformin 1000–1500 mg/day[a] (n = 50)	Percentage changes from baseline in HbA1c levels (secondary outcome)	4.03 [0.79, 7.27], p = 0.015
Trial 2	Takihata et al, 2013 [21]	2-arm RCT	Japanese adults with T2D inadequately controlled with metformin and/or SU	24 wk	Sitagliptin 50 mg/day (n = 58)	Pioglitazone 15 mg/day (n = 57)	Changes from baseline in HbA1c levels	−0.28 [−0.4, −0.16][b], p = 0.024
Trial 3	Onishi et al, 2013 [22]	2-arm RCT	Insulin-naïve, Japanese adults with T2D	26 wk	Once-daily injection of insulin degludec/insulin aspart (n = 147)	Once-daily injection of insulin glargine (n = 149)	Changes from baseline in HbA1c levels	−0.28 [−0.46, −0.10], p < 0.01

CI, confidence interval; *RCT*, randomized controlled trial; *SU*, sulfonylurea; *T2D*, type 2 diabetes; *wk*, weeks

[a]Initially administered 500 mg daily, followed by 1,000 mg daily after 2–4 weeks. The dose was increased to 1500 mg daily at 12 weeks in patients with inadequate glucose-lowering effects.

[b]The 95% CI was not presented in the manuscript; thus, we computed it using the mean differences, their standard deviations, and the number of patients based on the t-distribution.

clinical test values were taken has gaps for several months from baseline or the end of the study period, and this could affect the accuracy of the patient's background and outcome. In addition, doses of target drugs were not considered, which could cause misclassification in exposure definition. All criteria other than one exclusion criterion (Koshizaka et al.) or inclusion criterion (Onishi et al.) were considered. Ignoring this exclusion criterion or inclusion criterion could cause a difference in patient background compared to RCTs. Background differences between exposure and comparator groups were minimized by matching the propensity score.

2.4.2 Outcomes and Confounding Variables

The study outcome was either the change in HbA1c levels or the percentage change in HbA1c levels from baseline (Table 1). Baseline HbA1c levels were assessed 90 days before treatment initiation and post-treatment HbA1c levels were assessed 180–360 days after treatment initiation. The following potential confounders were measured using data at CED or in CED months: age, sex, duration of diabetes, and the Charlson Comorbidity Index [24].

2.5 Statistical Analyses

In each emulation, we implemented 1:1 propensity score (PS) nearest-neighbor matching using the above-listed potential confounders, with a caliper of 0.2 on the PS score scale, to balance the baseline patient characteristics between the groups. The baseline characteristics of the patients were summarized for both the pre- and post-matching populations, and standardized mean differences were calculated. An intention-to-treat (ITT) analysis was conducted, in which patients who started treatment were included and not censored regardless of discontinuation or change of treatment. The differences in study outcomes between the groups, their 95% confidence intervals (CIs), and p-values were calculated based on the *t*-distribution. Analyses were performed using SAS release 9.4 (SAS Institute, Cary, NC, USA).

2.5.1 Assessments of RCT–RWE Agreement

We used two binary metrics used in the RCT DUPLICATE Initiative to evaluate whether our RWE studies reproduced the same results as RCTs: (1) regulatory agreement and (2) estimate agreement [11]. The "regulatory agreement" refers to the ability of the RWE study to reproduce the direction and statistical significance of the findings of the RCT. The "estimate agreement" is met when the effect estimate obtained by the RWE study lies within the 95% CI for the effect estimate by the RCT. In the case of no 95% CI presented for the effect estimate in the RCT (Trial 2), we calculated the 95% CI using the estimates (mean differences), their standard deviations (SDs), and the number of patients based on the *t*-distribution.

2.5.2 Sensitivity Analyses

Several sensitivity analyses were conducted to explore the factors influencing agreement or disagreement between the results of RWE studies and RCTs. First, the effect estimates were calculated by modifying the time windows for the baseline and post-treatment HbA1c data. Second, summary statistics were calculated for the number and proportion of patients who discontinued RCT-allowed co-antidiabetic medications and those who had prescriptions of any other concomitant antidiabetic medications. When there was no prescription after the date of the previous prescription + supply days + 90 days (grace period), the medication was considered discontinued. Patients who discontinued the medication within 180 days from the CED were considered patients who discontinued the medication.

3 Results

3.1 Patient Characteristics

The baseline characteristics of the patients in our RWE studies are summarized in Table 3 along with the corresponding data in the RCTs. The number of patients in our emulation studies was equal to that in Trial 1 (48 vs 48 patients in the treatment group), more than that in Trial 2 (126 vs. 58 patients), and fewer than that in Trial 3 (61 vs. 147 patients). In all the RWE studies, the mean age of the patients was lower than that of the corresponding RCTs. Regarding sex distribution, the emulation studies for Trials 1 and 2 included fewer female patients than the RCTs, resulting in predominantly male patients. The mean baseline HbA1c levels in our emulation studies for Trials 1 and 3 were similar to those of the RCTs. However, in Trial 2 emulation, patients had higher mean ± SD HbA1c levels than in the RCT (treatment group, 7.8 ± 0.7 vs. 7.47 ± 0.66; comparator group, 7.9 ± 0.7 vs. 7.40 ± 0.61).

After PS matching, the standardized mean differences for each confounding factor were mostly within 0.25 in all emulations, indicating an acceptable balance of covariate distribution between the groups [25] (Table S2 of the ESM).

3.2 Results Between RWD Studies and RCTs

The between-group differences in outcome measurements in our emulations and the agreements between the RWE studies and RCTs are summarized in Table 4. In Trial 1 emulation, the percentage changes in HbA1c levels from

Table 2 Specification and emulation of a target trial

Protocol component	Target trial specification	Target trial emulation
Trial 1		
Eligible criteria	Individuals with the following criteria were eligible: diagnosed with type 2 diabetes according to the diabetes diagnostic criteria; aged 20–75 years; had received DPP-4i (sitagliptin 50 mg daily) for ≥12 weeks; had current HbA1c >7.0% and <10.0% and current BMI >22.0 kg/m^2; had estimated glomerular filtration rate >50.0 mL/min/1.73 m^2; and understood this study and provided written informed consent. Patients with the following criteria were excluded: type 1 diabetes; history of diabetic ketoacidosis, diabetic coma or pre-coma within 6 months prior to the date of consent; serious infections, surgery, or serious trauma requiring insulin therapy; moderate or high renal dysfunction (male serum creatinine level [Cre] ≥1.3 mg/dL, female Cre ≥1.2 mg/dL); hemodialysis treatment (including peritoneal haemodialysis); severe liver injury; history of serious vascular complications (stroke, myocardial infarction, and heart failure) requiring hospital admission; administration of glucose-lowering agents other than sitagliptin; women who were pregnant, lactating, possibly pregnant, or planning pregnancy; history of hypersensitivity to DPP-4 inhibitor, SGLT2 inhibitor, or metformin; presence or possibility of urinary tract infection or dehydration, positive urinary ketone body; history of lactic acidosis; excessive alcohol intake; history of fracture due to osteoporosis; CT examination conducted within 3 months prior to consent date; or study deemed inappropriate for the participant by their doctor	Same as for the target trial minus the exclusion criteria "administration of glucose-lowering agents other than sitagliptin". We also set a baseline period 90 days prior to the index date for covariate evaluation, and also requested the existence of HbA1c value in the period of − 180 to + 360 days of the end of the follow-up period
Treatment strategies	Patients in the ipragliflozin group received oral ipragliflozin 50 mg daily. Patients in the metformin group were initially administered 500 mg of metformin daily, and then 1000 mg daily after 2– 4 weeks	Same as for the target trial except for the dose (any dose was accepted)
Treatment assignment	Participants were randomly assigned to the ipragliflozin or metformin group in a 1:1 allocation	Randomization was emulated by adjusting for baseline covariates using propensity score matching (1:1)
Outcomes	The primary outcome was any change in the visceral fat area in 24 weeks between the two groups. CT imaging was performed before study drug administration and after 24 weeks. Two radiologists, who were masked to patients' clinical information and treatment assignment, centrally evaluated the CT images. Secondary outcomes included changes in HbA1c, body weight and BMI, waist circumference, fasting plasma glucose and insulin levels, homeostatic model assessment (HOMA)-beta, HOMA-R, total cholesterol, LDL-cholesterol, fasting triglycerides, and HDL-cholesterol, blood pressure, adiponectin, high sensitivity C-reactive protein (hs-CRP), subcutaneous and total fat area	Change in HbA1c
Follow-up	Starts at baseline and ends at loss to follow-up, discontinued intervention, or 24 weeks after baseline, whichever occurs first	Starts at baseline and ends at 24 weeks after baseline
Causal contrasts	Not specified	Intention-to-treat effect
Statistical analysis	Not specified	Intention-to-treat analysis

Table 2 (continued)

Protocol component	Target trial specification	Target trial emulation
Trial 2		
Eligible criteria	The inclusion criteria included type 2 diabetic men and women between the ages of 20–75 years whose diabetes had been inadequately controlled (HbA1c, 6.9–9.5%) with metformin and/or sulfonylurea. The following patients were excluded: (i) patients with a history of diabetic ketoacidosis or diabetic coma within 6months prior to study entry, (ii) patients with a history of cardiac failure, (iii) patients who underwent a surgical operation during the observation period of this study, (iv) patients with severe infection or severe trauma, (v) patients who were pregnant or lactating, (vi) patients with renal insufficiency [serum creatinine>132.6 μmol/L, estimated glomerular filtration rate (e-GFR) < 30 mL/min], (vii) patients with severe liver dysfunction, (viii) patients who had received insulin therapy, (ix) patients with a history of a hypersensitive reaction to sitagliptin or pioglitazone and (x) patients whowere judged as being inappropriate by the physicians in charge	Same as for the target trial We also set a baseline period 90 days prior to the index date for covariate evaluation, and also requested the existence of HbA1c value in the period of − 180 to +360 days of the end of the follow-up period
Treatment strategies	Sitagliptin group (sitagliptin, 50 mg/day) or a pioglitazone group (pioglitazone, 15 mg/day)	Same as for the target trial except for the dose (any dose was accepted)
Treatment assignment	All the subjects were randomly assigned in a 1:1 ratio to a sitagliptin group (sitagliptin, 50 mg/day) or a pioglitazone group (pioglitazone, 15 mg/day) by permuted block method using a central computer-based randomization and were followed up for 24 weeks	Randomization was emulated by adjusting for baseline covariates using propensity score matching (1:1)
Outcomes	The main outcome measure was the difference in the changes in the HbA1c levels from baseline value at 24weeks between these two groups, and the key secondary outcomes were the levels of fasting, plasma glucose (FPG) fasting insulin, inflammation mediators, *N*-terminal pro-B-type natriuretic peptide and markers of lipids, uric acid, liver function and renal function	Change in HbA1c
Follow-up	Starts at baseline and ends at loss to follow-up or 24 weeks after baseline, whichever occurs first	Starts at baseline and ends at 24 weeks after baseline
Causal contrasts	Intention-to-treat effect	Intention-to-treat effect
Statistical analysis	Intention-to-treat analysis	Intention-to-treat analysis

Table 2 (continued)

Protocol component	Target trial specification	Target trial emulation
Trial 3		
Eligible criteria	The trial recruited Japanese insulin-naıve adults with type 2 diabetes, aged > = 20 years with a glycosylated hemoglobin (HbA1c) concentration of 7–10% and a body mass index of < = 35 kg/m2. All subjects had been treated with > = 1 OAD(s) for >12 weeks and qualified for treatment intensification (HbA1c>7%). Subjects who had known or suspected allergy to any of the trial products, had serious heart disease, had impaired liver or renal function or had any disease or used any drugs that might interfere with glucose metabolism were excluded	Same as for the target trial minus the inclusion criteria "insulin-naive adults" We also set a baseline period 90 days prior to the index date for covariate evaluation, and also requested the existence of HbA1c value in the period of -180 to +360 days of the end of the follow-up period
Treatment strategies	The starting dose was 10U for both trial products. IDegAsp (70% IDeg and 30% IAsp, FlexPen®, 3 mL, 100 U/mL, Novo Nordisk A/S, Bagsværd, Denmark) was administered subcutaneously just prior to the largest meal of the day; the dosing time was chosen at the discretion of each subject and maintained throughout the trial. IGlar (Lantus®, SoloSTAR®, 3 mL, 100 U/mL, sanofiaventis, Frankfurt, Germany) was administered according to the approved labelling (either before breakfast or at bedtime at the discretion of each subject; the timing of the dosing was not to be changed during the trial). Insulin dose was titrated weekly based on the subject's mean prebreakfast selfmeasured plasma glucose (SMPG) measurement from the preceding 3 days (target: 3.9 to <5.0 mmol/l) according to a pre-defined titration algorithm	Same as for the target trial except for the dose (any dose was accepted)
Treatment assignment	Subjects were randomized (1:1) to treatment with either once-daily IDegAsp or once-daily IGlar, using a remote interactive voice/web response system for the randomization and stratification (by previous OAD treatment)	Randomization was emulated by adjusting for baseline covariates using propensity score matching (1:1)
Outcomes	The primary efficacy endpoint was the change from baseline in HbA1c after 26weeks of treatment. Other efficacy endpoints included change in laboratory-measured fasting plasma glucose (FPG) and nine-point SMPG profiles.	Change in HbA1c
Follow-up	Starts at baseline and ends at loss to follow-up or 26 weeks after baseline, whichever occurs first	Starts at baseline and ends at 26 weeks after baseline
Causal contrasts	Intention-to-treat effect	Intention-to-treat effect
Statistical analysis	Intention-to-treat analysis	Intention-to-treat analysis

[a]Data for RCTs are from Koshizaka et al. (2019) for Trial 1, Takihata et al. (2013) for Trial 2, and Onishi et al. (2013) for Trial 3.

Table 3 Baseline characteristics of patients in randomized controlled trials (RCTs) and emulated real-world evidence (RWE) studies

Variable	Group	Trial 1 (Koshizaka 2019)	Trial 2 (Takihata 2013)	Trial 3 (Onishi 2013)
N				
RCT[a]	Treatment	48 (49.0%)	58 (50.4%)	147 (49.6%)
	Comparator	50 (51.0%)	57 (49.6%)	149 (50.4%)
RWE study	Treatment	48 (50.0%)	126 (50.0%)	60 (50.0%)
	Comparator	48 (50.0%)	126 (50.0%)	60 (50.0%)
Female				
RCT[a]	Treatment	17 (35.4%)	22 (37.9%)	39 (26.5%)
	Comparator	22 (44.0%)	25 (43.9%)	34 (22.8%)
RWE study	Treatment	7 (14.6%)	15 (11.9%)	12 (20.0%)
	Comparator	4 (8.3%)	13 (10.3%)	12 (20.0%)
Age, years				
RCT[a]	Treatment	56.6 (11.9)	60.3 (7.5)	60.0 (10.0)
	Comparator	55.7 (12.2)	60.7 (9.5)	61.0 (9.6)
RWE study	Treatment	52.6 (7.4)	52.2 (7.8)	52.3 (7.8)
	Comparator	52.5 (7.2)	51.3 (7.0)	51.6 (7.8)
Duration of diabetes, years				
RCT[a]	Treatment	5.4 (4.6)	–	10.9 (7.3)
	Comparator	5.3 (4.8)	–	12.4 (8.6)
RWE study	Treatment	3.9 (3.4)	2.9 (2.2)	3.7 (2.6)
	Comparator	4.0 (3.3)	3.1 (2.1)	3.1 (2.7)
BMI (kg/m^2)				
RCT[a]	Treatment	27.55 (4.24)	24.6 (3.3)	25.2 (3.8)
	Comparator	28.83 (5.32)	25.8 (4.8)	25.0 (3.8)
RWE study	Treatment	28.7 (5.0)	27.5 (4.8)	25.7 (4.3)
	Comparator	26.8 (3.5)	28.2 (5.3)	26.1 (4.2)
HbA1c (%)				
RCT[a]	Treatment	7.95 (0.73)	7.47 (0.66)	8.3 (0.8)
	Comparator	8.12 (0.90)	7.40 (0.61)	8.5 (0.8)
RWE study	Treatment	8.1 (0.8)	7.8 (0.7)	8.4 (0.8)
	Comparator	8.0 (0.8)	7.9 (0.7)	8.2 (0.8)
Fasting glucose level[b] (mg/dl)				
RCT[a]	Treatment	159.9 (35.8)	143.9 (34.9)[c]	162.1 (28.8)[c]
	Comparator	166.1 (29.8)	142.0 (32.1)[c]	163.9 (34.2)[c]
RWE study	Treatment	168.5 (36.8)	157.6 (42.1)	158.0 (47.5)
	Comparator	154.6 (30.0)	154.0 (30.7)	155.7 (47.0)

BMI, body mass index; *RCT*, randomized controlled trial; *RWE*, real-world evidence

Data are presented as number (%) for *N* and female and mean (SD) for other variables

[a]Data for RCTs are from Koshizaka et al. (2019) for Trial 1, Takihata et al. (2013) for Trial 2, and Onishi et al. (2013) for Trial 3

[b]Data are fasting plasma glucose (FPG) for the RCT and fasting blood glucose (FBG) for the RWE study

[c]Unit was converted from mmol/L to mg/dL by dividing it by 0.05551

baseline were larger in the treatment group than in the comparator group (difference [treatment – comparator] −6.21, 95% CI −11.01 to −1.40; $p = 0.012$). This result was in the opposite direction to that of the RCT. Similarly, emulations of Trials 2 and 3 did not yield the same results as those of the RCTs. Changes in HbA1c levels from baseline were larger in the treatment group than in the comparator group in Trial 2 (difference −0.01; 95% CI −0.25 to 0.23; p = 0.926), and smaller in Trial 3 (difference 0.46; 95% CI −0.01 to 0.94; $p = 0.056$). In all three emulations, neither regulatory nor estimate agreement was achieved.

Table 4. Effect estimates and RCT–RWE agreements

Study	Outcome measurement	RCT[a]				RWE study				RCT–RWE agreement	
		N (TRT/COMP)	Difference between groups (TRT−COMP)			*N* (TRT/COMP)	Difference between groups (TRT−COMP)			Regulatory agreement	Estimate agreement
			Mean	95% CI	*p* value		Mean	95% CI	*p* value		
Trial 1	Percentage changes in HbA1c levels (%)	48/50	4.03	0.79, 7.27	0.015	48/48	−6.21	−11.01, −1.40	0.012	Not agreed	Not agreed
Trial 2	Changes in HbA1c levels (%)	58/57	−0.28	−0.4, −0.16	0.024	126/126	−0.01	−0.25, 0.23	0.926	Not agreed	Not agreed
Trial 3	Changes in HbA1c levels (%)	147/149	−0.28	−0.46, −0.10	<0.01	60/60	0.46	−0.01, 0.94	0.056	Not agreed	Not agreed

CI, confidence interval; *COMP*, comparator group; *RCT*, randomized controlled trial; *RWE*, real-world evidence; *TRT*, treatment group

[a]Data for RCTs are from Koshizaka et al. (2019) for Trial 1, Takihata et al. (2013) for Trial 2, and Onishi et al. (2013) for Trial 3.

3.3 Results of Sensitivity Analyses

We modified the time windows for baseline HbA1c data (180 days and 60 days before treatment initiation, instead of 90 days) and post-treatment HbA1c assessments (180–330 days after treatment initiation, instead of 180–360 days). However, these modifications did not alter the conclusions (Table S3 of the ESM).

The proportion of patients who discontinued the RCT-allowed concomitant antidiabetics (i.e., DPP-4i in Trial 1, metformin or sulfonylurea in Trial 2, and any oral antidiabetics in Trial 3) during the follow-up, which was not considered in the primary analysis, was less than 15% in any emulation (Table S4 of the ESM). However, a substantial proportion of patients used antidiabetics other than the group's treatment and the RCT-allowed co-medications in both the treatment and comparator groups: Trial 1 emulation, 62.5% and 64.1%; Trial 2, 39.7% and 81.0%; and Trial 3, 16.7% and 61.7%, respectively (Table S4 of the ESM).

4 Discussion

This was the first attempt to replicate RCTs with a Japanese database of claims and health checkup data to examine whether RWE studies can produce the same conclusions as RCTs if carefully designed and analyzed. Of the 13 candidate RCTs evaluating the treatment effects of diabetic medications on HbA1c levels, only three were feasible for replication using this RWD source, primarily due to a lack of necessary data. This major challenge limits opportunities for RWE studies, as observed in previous studies [13, 26]. In all three emulation studies with RWD, the obtained results did not meet either "regulatory agreement" or "estimate agreement" with the results from RCTs, demonstrating that this database was not the best fit for these research questions.

As the JMDC database contains health checkup results in addition to claims data, we expected that it could be utilized for effectiveness evaluation using laboratory data as outcomes. However, many patients in the database lacked clinical data to define the population and outcomes, resulting in only a few replicable RCTs. One reason for the lack of HbA1c data was missing values; health checkup results were not collected from all health insurance societies contributing to this database [19]. Another reason is that people in Japan usually undergo health check-ups once a year. These infrequent data further reduced the number of patients with HbA1c data within specific time windows. If laboratory data of frequent intervals or precise timing are essential variables in the study, an RWE study would not be feasible using yearly health checkup data.

As suggested in previous studies, the discrepancies in results between RWE studies and RCTs can arise from differences in design, such as the study population, treatment patterns, and outcome measurement [13, 27] in addition to the lack of randomization. For example, a previous study suggested that heterogeneity in patient characteristics may lead to different results between the emulated RWE study and RCTs, and in that case, evaluation of the agreement between them is not feasible [28]. In our study, patient characteristics, such as age, sex, duration of diabetes, BMI, and baseline HbA1c levels, differed between emulations and RCTs. The authors of Trial 2 argued that BMI might affect the effectiveness of sitagliptin [21]. The mean BMI in our emulation study was higher than that in the RCT, which might be partly responsible for the different conclusions. Examining the clinical reasons behind the RWE–RCT differences was beyond the scope of this study; therefore, we will not go elaborate on these such details. Moreover, these differences in study populations do not necessarily indicate the drawbacks of RWE studies; instead, they are essential to fill the efficacy–effectiveness gap [29]. Nevertheless, researchers should bear in mind that RWE studies can result in populations that are different from RCTs, even if rigorously designed.

Some of these differences were probably introduced because we could not precisely mirror some inclusion/exclusion criteria due to the lack of data and other constraints of the data source, as in previous attempts [13, 27]. For example, we had to loosen the condition of antidiabetic medications to secure the number of patients, which undoubtedly diverged patient selection and treatment patterns in the RWE studies. Indeed, most patients in our emulations, used other antidiabetics in addition to the study treatment. This is a typical example of the difficulty of tightly controlling treatment settings in RWE studies. As mentioned earlier, such data reflecting actual clinical practice are essential for filling the efficacy–effectiveness gap [30]. However, this complexity of RWE studies is the very thing that complicates the interpretations of the study results, posing hurdles for their use as valid evidence about the treatment's effectiveness [2]. Thus, for an RWE study with such an aim, not as a supplement to RCTs, it would still be crucial to simplify the settings as much as possible. In this sense, it is important to understand that RWE studies have limited opportunities depending on the research questions, as demonstrated in this study.

Furthermore, the extended time windows for HbA1c data must also have introduced differences between the RWE studies and RCTs. We had to set broad time windows because patients usually had only one HbA1c data point yearly. Therefore, their HbA1c data would not have adequately reflected the glycemic condition at the precise timing of treatment initiation or the end of follow-up. This impreciseness is a significant design limitation in our emulations. Our sensitivity analyses using different time windows primarily resulted in the same trends as the primary analysis, suggesting that these time windows had no significant impact on the outcomes. However, the modified time window (i.e., 180–330 days after treatment initiation) was still a long way off from the end of follow-up; thus, the results may have changed if data exactly at the end of follow-up were analyzed.

In this study, we found that this RWD source was not feasible for evaluating the effectiveness of diabetic medications on HbA1c levels because of the lack of data critical for designing these studies. However, the disagreement in results between RWE studies and original RCTs, as in the present study and similar attempts [13, 27], does not necessarily indicate the low quality of the data sources or analyses. For example, this database may still be useful for evaluating yearly changes in blood test data or evaluating an outcome that can be defined by a diagnostic record in claims data with high accuracy. Instead, understanding whether a particular RWD source fits the study of interest is a significant finding in RCT replication exercises. Accumulating such knowledge will help to further understand when and how we can implement a high-quality RWE study to produce a valid conclusion. Therefore, RCT replication exercises such as ours should be performed more vigorously in various clinical settings and RWD sources.

This study evaluated the feasibility of claims and health checkup data for RCT replication. Previous RCT replication attempts used data sources such as claims data [15, 27], electronic health records (EHRs) [14], and registry data [13], but not health checkup data. Thus, the findings of this study will add new information to the existing knowledge from replication exercises. However, this study also has limitations. First, we emulated only three RCTs. Thus, the generalizability of our findings may be limited, and similar replication exercises may yield different results. Second, despite using a large database in this study, the number of subjects in emulated RWEs was relatively small. Therefore, the characteristics of the sampled population may be biased, which also limits the generalizability of the results. Furthermore, the small sample size in our emulated RWE study is considered to increase the variability of estimates and reduce the statistical power. In general, the RWE study requires at least as great a sample size as calculated in the corresponding RCT study. However, in one emulation of our study, there were fewer study subjects than the corresponding RCT, which made our conclusion difficult due to random error. Third, to get comparable results, the distribution of patient characteristics in the emulated RWE study should have been matched with the RCTs as recommended in the previous study [28]. However, it was not implemented in our study due to the small number of study subjects. Forth, since only subjects with measured outcome variable were included, there was the possibility

of selection bias. Fifth, an ITT analysis was implemented in this study. Since adherence to medications are generally poor in RWD relative to RCTs, exposure misclassification is likely to have occurred. However, although a per-protocol approach may reduce this type of misclassification, there is concern that informative censoring is important. Therefore, we adopted the ITT analysis because it is straightforward. Sixth, our RWE studies did not precisely mimic various study elements, including eligibility criteria and outcome measures, primarily because of the limitations of the data sources. A different data source, such as EHR, might have replicated these RCTs. Thus, it should be noted that our results do not necessarily deny the feasibility of all RWE studies for these research questions. Furthermore, the data source used in our study could also be used to replicate RCTs with outcomes that can be emulated using a diagnostic record.

5 Conclusion

In conclusion, our RWE studies using a Japanese claims and health checkup database did not reproduce the same conclusions as the RCTs that evaluated the treatment effects of diabetic medications on HbA1c levels in patients with type 2 diabetes in Japan. The results of this RCT replication attempt suggested that this particular RWD source may not be suitable for evaluating treatment effects using laboratory data as the study outcomes. We expect that further RCT replication attempts should be conducted in various clinical areas using Japanese RWD to accumulate knowledge on the opportunities and limitations of RWE studies in Japan.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40801-023-00353-7.

Acknowledgements Clinical Study Support Inc. provided medical writing support.

Author contributions RW, TH, and TK contributed to study conception and design. TH conducted the statistical analyses. RW drafted the manuscript. All authors contributed to data interpretation and critically revised the manuscript for important intellectual content. All authors read and approved the final version of the manuscript for submission.

Funding This study was conducted as part of a joint research course (Real-World Evidence and Data Assessment) between Juntendo University and Shin Nippon Biomedical Laboratories, Ltd.

Data Availability The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest RW and TH are employees of Clinical Study Support Inc. TK is on the boards of Clinical Study Support Inc. RK has no conflict of interests to declare.

Ethics Approval This study was approved by the ethics committee of Juntendo University (approval number: E21-0284). This study was conducted in accordance with the Declaration of Helsinki.

Consent to Participate This observational study used only secondary fully anonymized data. An observational study that exclusively used anonymized data is outside the scope of the ethical guidelines for medical and health research involving human subjects in Japan. Therefore, informed consent was not obtained for this study.

References

1. U.S. Food and Drug Administration. Framework for FDA's real-world evidence program. 2018. https://www.fda.gov/media/120060/download. Accessed 7 Sep 2022.
2. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? Clin Pharmacol Ther. 2017;102:924–33.
3. Schneeweiss S, Patorno E. Conducting real-world evidence studies on the clinical outcomes of diabetes treatments. Endocr Rev. 2021;42:658–90.
4. Kim HS, Lee S, Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. J Korean Med Sci. 2018. https://doi.org/10.3346/jkms.2018.33.e213.
5. Nishioka K, Makimura T, Ishiguro A, Nonaka T, Yamaguchi M, Uyama Y. Evolving acceptance and use of RWE for regulatory decision making on the benefit/risk assessment of a drug in Japan. Clin Pharmacol Ther. 2022;111:35–43.
6. Schneeweiss S. Improving therapeutic effectiveness and safety through big healthcare data. Clin Pharmacol Ther. 2016;99:262–5.
7. Sheffield KM, Dreyer NA, Murray JF, Faries DE, Klopchin MN. Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions. J Comp Eff Res. 2020;9:1043–50.
8. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: the changing landscape of the regulatory approval process. Pharmacoepidemiol Drug Saf. 2020;29:1201–12.
9. Beaulieu-Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, et al. Examining the use of real-world evidence in the regulatory process. Clin Pharmacol Ther. 2020;107:843–52.

10. Hiramatsu K, Barrett A, Miyata Y, PhRMA Japan Medical Affairs Committee Working Group. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs Real World Outcomes. 2021;8:459-80.
11. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, et al. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. Clin Pharmacol Ther. 2020;107:817–26.
12. RCT DUPLICATE [website]. Project. https://www.rctduplicate.org/projects.html. Accessed 10 Aug 2022.
13. Jemielita T, Widman L, Fox C, Salomonsson S, Liaw KL, Pettersson A. Replication of oncology randomized trial results using Swedish registry real world-data: a feasibility study. Clin Pharmacol Ther. 2021;110:1613–21.
14. Wallach JD, Deng Y, McCoy RG, Dhruva SS, Herrin J, Berkowitz A, et al. Real-world cardiovascular outcomes associated with degarelix vs leuprolide for prostate cancer treatment. JAMA Netw Open. 2021. https://doi.org/10.1001/jamanetworkopen.2021.30587.
15. Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial: Cardiovascular safety of linagliptin versus glimepiride. Diabetes Care. 2019;42:2204–10.
16. Laurent T, Simeone J, Kuwatsuru R, Hirano T, Graham S, Wakabayashi R, et al. Context and considerations for use of two Japanese real-world databases in Japan: Medical Data Vision and Japanese Medical Data Center. Drugs Real World Outcomes. 2022;9:175–87.
17. Ministry of Health Labour and Welfare. A Basic Direction for Comprehensive Implementation of National Health Promotion. 2012. https://www.mhlw.go.jp/file/06-Seisakujouhou-10900000-Kenkoukyoku/0000047330.pdf. Accessed 5 Sep 2022.
18. International Diabetes Federation. Diabetes in Western Pacific–2021. https://diabetesatlas.org/idfawp/resource-files/2021/11/IDF-Atlas-Factsheet-2021_WP.pdf. Accessed 5 Sep 2022.
19. Nagai K, Tanaka T, Kodaira N, Kimura S, Takahashi Y, Nakayama T. Data resource profile: JMDC claims database sourced from health insurance societies. J Gen Fam Med. 2021;22:118–27.
20. Koshizaka M, Ishikawa K, Ishibashi R, Maezawa Y, Sakamoto K, Uchida D, et al. Comparing the effects of ipragliflozin versus metformin on visceral fat reduction and metabolic dysfunction in Japanese patients with type 2 diabetes treated with sitagliptin: a prospective, multicentre, open-label, blinded-endpoint, randomized controlled study (PRIME-V study). Diabetes Obes Metab. 2019;21:1990–5.
21. Takihata M, Nakamura A, Tajima K, Inazumi T, Komatsu Y, Tamura H, et al. Comparative study of sitagliptin with pioglitazone in Japanese type 2 diabetic patients: the COMPASS randomized controlled trial. Diabetes Obes Metab. 2013;15:455–62.
22. Onishi Y, Ono Y, Rabol R, Endahl L, Nakamura S. Superior glycaemic control with once-daily insulin degludec/insulin aspart versus insulin glargine in Japanese adults with type 2 diabetes inadequately controlled with oral drugs: a randomized, controlled phase 3 trial. Diabetes Obes Metab. 2013;15:826–32.
23. JMDC Inc [website]. JMDC Real World. Publication records. https://www.phm-jmdc.com/publications. Accessed 5 Sep 2022.
24. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43:1130–9.
25. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25:1–21.
26. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of using real-world data to replicate clinical trial evidence. JAMA Netw Open. 2019. https://doi.org/10.1001/jamanetworkopen.2019.12869.
27. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. Circulation. 2021;143:1002–13.
28. Lin LA, Zhang Y, Straus W, Wang W. Integrative analysis of randomized clinical trial and observational study data to inform post-marketing safety decision-making. Ther Innov Regul Sci. 2022;56:423–32.
29. Nordon C, Karcher H, Groenwold RH, Ankarfeldt MZ, Pichler F, Chevrou-Severac H, et al. The "efficacy-effectiveness gap": Historical background and current conceptualization. Value Health. 2016;19:75–81.
30. Thompson D. Replication of randomized, controlled trials using real-world data: What could go wrong? Value Health. 2021;24:112–5.

Authors and Affiliations

Ryozo Wakabayashi[1,2] · Takahiro Hirano[1,2] · Tadashi Koga[1,2] · Ryohei Kuwatsuru[2,3]

✉ Ryozo Wakabayashi
wakabayashi.ryozo@gmail.com

[1] Clinical Study Support, Inc., Nagoya, Japan

[2] Real-World Evidence and Data Assessment (READS), Graduate School of Medicine, Juntendo University, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan

[3] Department of Radiology, School of Medicine, Juntendo University, Tokyo, Japan