Kurki, Samu; Halla-aho, Viivi; Haussmann, Manuel; Lähdesmäki, Harri; Leinonen, Jussi V.;
Koskinen, Miika

A comparative study of clinical trial and real-world data in patients with diabetic kidney
disease

scientific reports

OPEN

A comparative study of clinical trial and real-world data in patients with diabetic kidney disease

Samu Kurki[1,4✉], Viivi Halla-aho[2,4], Manuel Haussmann[3], Harri Lähdesmäki[3], Jussi V. Leinonen[1] & Miika Koskinen[2]

A growing body of research is focusing on real-world data (RWD) to supplement or replace randomized controlled trials (RCTs). However, due to the disparities in data generation mechanisms, differences are likely and necessitate scrutiny to validate the merging of these datasets. We compared the characteristics of RCT data from 5734 diabetic kidney disease patients with corresponding RWD from electronic health records (EHRs) of 23,523 patients. Demographics, diagnoses, medications, laboratory measurements, and vital signs were analyzed using visualization, statistical comparison, and cluster analysis. RCT and RWD sets exhibited significant differences in prevalence, longitudinality, completeness, and sampling density. The cluster analysis revealed distinct patient subgroups within both RCT and RWD sets, as well as clusters containing patients from both sets. We stress the importance of validation to verify the feasibility of combining RCT and RWD, for instance, in building an external control arm. Our results highlight general differences between RCT and RWD sets, which should be considered during the planning stages of an RCT-RWD study. If they are, RWD has the potential to enrich RCT data by providing first-hand baseline data, filling in missing data or by subgrouping or matching individuals, which calls for advanced methods to mitigate the differences between datasets.

Clinical real-world data (RWD) has garnered increasing interest for its use alongside clinical trial protocols in generating evidence in medical research. Several studies have explored the utilization of RWD in conducting clinical trials[1–4], providing external control groups for single-arm studies[5,6], and complementing the control group in randomized controlled trials (RCTs)[7–9]. The rationale for using RWD to augment RCTs relies on the assumption that RWD and RCT datasets are comparable despite potential biases. However, data generation mechanisms differ substantially between RWD and RCT, and systematic differences are probable. As a result, verifying the compatibility between datasets becomes a crucial component of data preprocessing to ensure the study's feasibility.

Compared to clinical trial data, RWD's quality can greatly fluctuate depending on its purpose and the specific dataset. Characteristics such as accuracy, completeness, and sampling intervals can vary among covariates, patients, and healthcare providers[10,11]. Certain clinical measures, such as blood pressure and weight, are often available for research, but variables like outpatient medication exposure might need to be inferred indirectly from prescriptions and are potentially overestimated if prescriptions were left unused[11]. Choices in study design, like index date or lookback window, can influence prevalence and incidence estimates based on claims and electronic health record (EHR) databases[12]. These factors may not be as clearly defined in RWD as in RCTs, potentially affecting the temporal alignment of the datasets. There might also be other underlying time-related biases which complicate, for instance, the identification of event onset or exposure to treatment[13]. For example, a delay between disease initiation and detection can introduce temporal variation and bias into the data[13].

Apart from inclusion and exclusion criteria, the discrepancy between datasets can be alleviated through the application of appropriate validation criteria for the relevance, reliability, and quality of the RWD[14] and by particularly focusing on confounders related to exposure and outcome[15]. Several methods for selecting control patients have been suggested, such as cardinality matching of individuals[15] based on propensity scoring[16], which has become the preferred method for adjusting group differences and reducing confounding. However, selection criteria and computational advancements are merely parts of the solution and should not be relied upon unconditionally.

[1]Bayer Oy, Tuulikuja 2, 02100 Espoo, Finland. [2]HUS Helsinki University Hospital, Helsinki, Finland. [3]Department of Computer Science, Aalto University, Espoo, Finland. [4]These authors contributed equally: Samu Kurki and Viivi Halla-aho. ✉email: samu.kurki@bayer.com

In this study, we compared baseline data from a completed clinical trial on chronic kidney disease outcomes in individuals with type 2 diabetes to that from electronic health records (EHRs) in order to characterize their similarities and differences. Our focus was on five common data types: demographics, diagnoses, medications, laboratory measurements, and vital signs. We evaluated temporal aspects such as extent and sampling density, along with data completeness or missingness. In addition to comparisons between whole datasets, through statistical and cluster analysis, we demonstrate a partial overlap of RWD and RCT data.

Methods
Data
We used RCT data from the completed FIDELIO-DKD trial (Bayer, NCT02540993) to study the effect of finerenone on chronic kidney disease outcomes in type 2 diabetes in adult patients (≥ 18 years)[17]. Patients included in the trial had diabetes and fulfilled the following criteria for kidney disease at the time of randomization: a urinary albumin-to-creatinine ratio (UACR) of 30–300 (mg/g), estimated glomerular filtration rate (eGFR) of 25–60 (ml/min/1.73m^2), or a UACR of 300–5000 and eGFR of 25–75 (ml/min/1.73m^2) along with diabetic retinopathy (see[17] for further details). The data were pseudonymized, and we used only baseline data prior to randomization, including demographics, vital signs, diagnosis history, laboratory measurements, and concomitant medications. We obtained internal approval for secondary research use of the trial data.

To be included in the RWD set, patients needed to have chronic kidney disease as defined by ICD-10 codes N18, N19, or I12, or eGFR < 45 ml/min/1.73m^2 at some time point in their EHRs. Thus, the patients had moderate to severe impairment in kidney function based on the Kidney Disease Improving Global Outcomes (KDIGO) guidelines. UACR criteria were not used in the RWD set due to its low availability. Type 2 diabetes mellitus was also required, defined either by diagnosis code E11, the use of diabetes medication (ATC class A10), or a glycated hemoglobin (H-HbA1c) measurement ≥ 48 mmol/mol.

The RWD set contained data from patients who were first diagnosed with either chronic kidney disease or type 2 diabetes mellitus as adults. We assessed this post hoc using data from medications, diagnoses, and laboratory tests. We extracted these data along with demographics and vital signs from electronic healthcare records of HUS Helsinki University Hospital, Finland, covering a ten-year period from 2012 to 2021.

The index date in the RCT data was the date of randomization, while in the RWD, it was the date when the chronic kidney disease inclusion criteria were met.

Data structuring
The RCT data were structured. RWD were also structured, with the exception of smoking status, some medications, and New York Heart Association (NYHA) classes. We extracted these data from clinical documents using text mining and added them to the dataset with corresponding timestamps.

Data harmonization
We used SNOMED coding to harmonize the different nomenclatures of RCT and RWD. The mapping was straightforward for demographics, medications, laboratory measurements, and vital signs. For diagnoses, 95.9% of MedDRA codes in RCT and 94.1% of ICD-10 codes in RWD were successfully mapped to SNOMED coding. Unmapped diagnoses codes, which were primarily in the Z-category indicating factors affecting health status or contact with health services, were not considered critical and were excluded from subsequent analyses. For diagnoses, we used the latest version of the mapping table from OHDSI Athena[18]. We performed standard unit conversions between RCT and RWD laboratory values for laboratory measurements.

Analytics environment
We stored and processed the data on the HUS Acamedic cloud-based data analytics platform[19]. This platform enables high-performance scientific computing and can be scaled as necessary. The platform meets both European and national regulations (General Data Protection Regulation, Finlex 552/2019) for processing sensitive health and social data, it has a valid security certification, and is supervised by the National Supervisory Authority for Welfare and Health (Valvira).

Statistical methods
We compared the prevalence of diagnoses and medications between RCT and RWD sets and reported counts and proportions by category. We compared the data longitudinality and sampling density aspects, namely the length of pre-index time, events per year, unique codes per patient, and interval between events, between RCT and RWD sets as continuous variables. These were reported with medians and interquartile ranges (IQR). We used chi-squared tests for group comparisons with categorical variables and the Kruskal–Wallis H test for continuous data. Moreover, we computed odds ratios to evaluate similarity of features within different clusters of patients. We performed the analyses with Python 3.8.10.0 using pandas (version 1.1.5)[20] and SciPy (version 1.5.3)[21] libraries.

Cluster analysis of diagnoses and medications
To further examine and visually represent the differences and overlaps between the RCT and RWD sets, we performed separate cluster analyses for medication and diagnosis datasets, which were formed by merging the RCT and RWD sets. These analyses were similar to those outlined in[22]. For this purpose, we mapped diagnoses in the RCT data to International Classification of Diseases version 10 codes (ICD-10) as detailed in the Results section. Following the method in[22], we limited our analysis to ICD-10 diagnosis categories A to N with a precision of three characters. Hence, our focus was on disorders while excluding codes related to pregnancy, external causes,

malformations, and contacts with health services. For medication data, we selected the first four characters of the Anatomical Therapeutical Chemical (ATC) codes. In both data sets, a specific code had to have at least 1% prevalence to be included in the analysis. Consequently, the data sets contained 65 covariates for diagnoses and 84 covariates for medications.

Following a two-step approach for clustering[22], we initially trained a variational autoencoder (VAE) model[23,24] using Keras[25] (version 2.3.1). This model projected binary diagnosis vectors into a two-dimensional latent space[23,24]. Then we clustered the projected vectors using the HDBSCAN algorithm[26].

For the encoder and decoder components of the VAE models, we utilized fully connected multilayer perceptrons (MLPs) with a single hidden layer, with either a hyperbolic tangent (tanh) or a rectified linear unit (ReLU)[27] as the activation function. Since the purpose of the cluster analysis was to visually distinguish the data sets and their differences, we selected the activation function that produced visually distinct subgroups. The VAE model was trained using the evidence lower bound objective, which comprises a reconstruction loss and a regularizer on the latent space. We divided the data into a training set (90% of the data) and a validation set (10% of the data), with the validation data used to choose a hyperparameter for the number of gradient descent steps for training. Eventually, we mapped both sets to the latent space for the clustering step.

We used the HDBSCAN algorithm[26] (version 0.8.29) to extract clusters for subsequent description and interpretation of the identified subgroups. HDBSCAN is a density-based clustering algorithm that groups similar data points together while also identifying outliers. It constructs a hierarchy of clusters by considering local density variations and connectivity, allowing for the automatic determination of the number of clusters. We chose HDBSCAN due to its ability to automatically determine the number of clusters, handle varying cluster shapes and sizes, as well as noise. For diagnosis data, we used the HDBSCAN parameters min_cluster_size = 220 and min_samples = 1, and for medications data, we used min_cluster_size = 200 and min_samples = 5. To visualize the results of the cluster analysis, we used NumPy[28] (version 1.21.6) to compute two-dimensional histograms, SciPy[21] (version 1.5.3) for kernel density estimation, and Matplotlib[29] (version 3.2.1) for plotting. To preserve patient anonymity, we avoided presenting the locations of individual data points and instead utilized histogram and density-based approaches for visualization.

Ethical aspects

According to Finnish legislation (Act on the Secondary Use of Health and Social Data (552/2019) by the Ministry of Social Affairs and Health), the approval of an ethical committee or informed consent is not required for non-interventional, observational retrospective registry studies. The study was conducted in accordance with the Declaration of Helsinki and the General Data Protection Regulation (GDPR). HUS Helsinki University Hospital approved the study (permission HUS/230/2022). The original RCT data collection was based on informed consent from patients participating in the FIDELIO-DKD clinical trial (ClinicalTrials.gov identifier NCT02540993).

Results

Qualitative comparisons

We analyzed the pre-index baseline data of 23,523 RWD and 5,734 RCT patients, all of whom had both chronic kidney disease and type 2 diabetes. Harmonizing to common nomenclature for both RCT and RWD sets was straightforward, but we noted considerable qualitative differences in data generation, temporality, and completeness, as summarized in Table 1. For instance, in the RCT data, diagnoses and concomitant medications based on case report forms (CRFs) collected by investigators spanned up to 50 years pre-index. Conversely, in the RWD, all diagnoses in EHRs with precise dates were available up to 10 years pre-index, constrained by the research permit. Laboratory measurements and vital signs were only available near index in RCT data and up to 10 years pre-index in RWD.

Quantitative comparisons

Statistical analysis revealed notable differences in completeness, particularly in medications and diagnoses. After harmonizing the different nomenclatures of RCT and RWD, we used ATC and ICD-10 code classes for easier interpretation of the results. In concomitant medications (Fig. 1A), the most significant difference in prevalence was observed in anti-infectives for systemic use (class J, RCT 8.6%, RWD 66.0%, $P < 0.001$). In diagnosis history (Fig. 1B), the largest difference in prevalence is seen in class R, which refers to symptoms, signs, and abnormal clinical and laboratory findings (RCT 14.3%, RWD 57.1%, $P < 0.001$). Although fulfilling the inclusion criteria, not all patients in the RWD had both inclusion diagnoses of type 2 diabetes and chronic kidney disease, which belong to classes E (endocrine, nutritional, and metabolic diseases, RCT 100.0%, RWD 50.7%, $P < 0.001$) and N (diseases of the genitourinary system, RCT 100.0%, RWD 49.9%, $P < 0.001$), respectively.

As demographics, laboratory measurements, and vital signs in the RCT data were only available near the index time, we further analyzed the temporal characteristics of concomitant medications and diagnoses (Table 2). We defined event as a patient encounter in either the outpatient or inpatient setting. The pre-index time (in years) from the first event to the index date was significantly longer in RCT data for both medications (median 7.7 vs. 4.9, $P < 0.001$) and diagnoses (median 20.3 vs. 5.2, $P < 0.001$). We also assessed the time interval between events and the sampling density, defined as the number of events per year. The number of events per year and the number of unique codes were significantly larger in RWD compared to trial data for both medications and diagnoses. Consequently, the time interval between consecutive events was significantly shorter in RWD for both medications and diagnoses compared to RCT. Additionally, we observed a significant difference between RWD and RCT cohorts in age at index (median 76 years in RWD, median 66 in RCT, $P < 0.001$) and sex male/female ratio (RWD 47/53, RCT 70/30, $P < 0.001$).

Property	Data				
	Demographics (DM)	**Diagnosis history (DH)**	**Laboratory measurements (LB)**	**Vital signs (VS)**	**Concomitant medications (CM)**
Data generation	RCT: Measured from patients at screening. RWD: As measured in healthcare	RCT: CRF and discretion of a physician. RWD: All diagnoses during the years as in an institutional database	RCT: Measured from patients at screening. RWD: As measured in healthcare	RCT: Measured from patients at screening. RWD: As measured in healthcare	RCT: CRF and discretion of a physician. RWD: Prescriptions and inpatient drugs administered. Mentions in text
Longitudinality	RCT: Values available at index date (time of randomization). RWD: Time series. Index date was defined to calculate age and other variables at index	RCT: Time series, up to 50 years of historical data. No exact dates. RWD: Time series up to 10 years of historical data, limited by research permit. Exact dates	RCT: Only available near index date. RWD: Time series up to 10 years of historical data available, limited by research permit. Exact dates	RCT: Only available near index date. RWD: Time series up to 10 years of historical data available, limited by research permit. Exact dates	RCT: Time series, up to 50 years of historical data. No exact dates.RWD: Time series up to 10 years of historical data, limited by research permit. Exact dates
Completeness	RCT: Values available at index date. RWD: Obtained from time series of structured data except smoking status and NYHA class which were extracted from structured and text sources	RCT: Some diagnoses underrepresented, such as infections (Fig. 1).RWD: Some diagnoses underrepresented, such as genitourinary diseases (Fig. 1)	RCT: All values available from a given set of measurements. RWD: Missing values with respect to trial design, availability depends on healthcare practices	RCT: All values available from a given set of measurements. RWD: Missing values with respect to trial design, availability depends on healthcare practices	RCT: Selected medications. Some medications underrepresented in the data, such as anti-infectives (Fig. 1). RWD: All prescriptions and administered inpatient medications in institutional database. Selected medications extracted from text
Nomenclature	Commonly used demographic variables, such as age, gender, and smoking	RCT: MedDRA codes. RWD: ICD-10 codes	RCT: MedDRA codes. RWD: Logical Observation Identifiers Names and Codes (LOINC)	Commonly used vital sign variables, such as BMI, blood pressure and heart rate	ATC codes both in RCT and RWD structured data. Also, medication labels in text
Harmonization	Same nomenclature	RCT: 95.9% of MedDRA codes successfully mapped to SNOMED. RWD: 94.1% of ICD-10 codes successfully mapped to SNOMED	Straightforward mapping between laboratory measurements. Standard unit conversions were performed between RCT and RWD laboratory values	Same nomenclature	Same nomenclature

Table 1. Qualitative observations between the five medical domains in the RCT and RWD data sets.
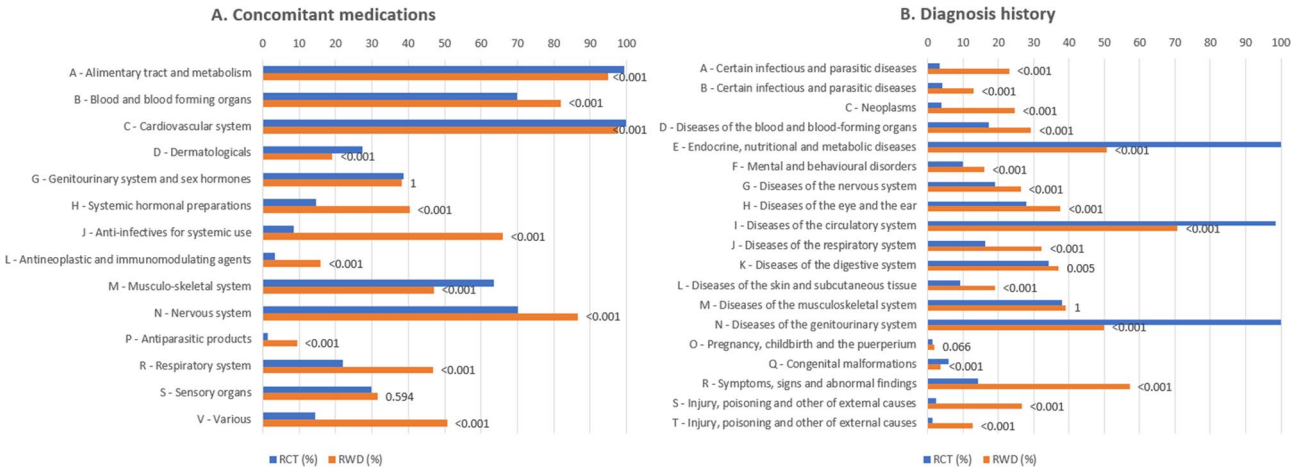

Figure 1. Comparison of proportion (%) of RCT ($N = 5\ 734$) and RWD ($N = 23\ 523$) patients with different medication/diagnosis types recorded, each pair of RCT and RWD bars corresponding to an ATC class (**A**) or ICD-10 class (**B**). Value labels above bars correspond to P values calculated with chi-squared tests. Presented P values were adjusted with Bonferroni correction to address the issue of multiple comparisons.

Cluster analysis of diagnoses and medications

To elucidate the overlap between RWD and RCT cohorts, we conducted cluster analyses of diagnoses and medications, integrating the RWD and RCT datasets for this purpose. Figures 2 and 3 graphically represent the outcomes of these analyses. In both the diagnosis and medication datasets, the RWD and RCT cohorts emerged as discernable subgroups in the latent space of the Variational Autoencoder (VAE) model. Eleven clusters were discerned from the diagnosis data, and seven from the medication data. For the diagnosis dataset, 5133 patients (18%) did not align with any cluster; a comparable figure for the medication dataset was 5487 patients (19%). Detailed overviews of each cluster can be found in Tables 3 and 4, and in the supplement.

From the diagnosis dataset, three clusters (0, 1, and 3) mainly consisted of RCT patients, while eight clusters (2, 4–10) primarily consisted of RWD patients. Cluster 4 demonstrated the most substantial overlap, with 22% of its patients from the RCT cohort and 78% from the RWD cohort. Cluster 3 also exhibited significant overlap, with 89% of its data derived from the RCT cohort and 11% from the RWD cohort. This overlap signifies the

Concomitant medications					Diagnosis history				
Variable	Quartile	RCT	RWD	P value*	Variable	Quartile	RCT	RWD	P value*
Pre-index time (Years)	25%	3,6	2,4	< 0.001	Pre-index time (Years)	25%	14,9	2,7	< 0.001
	Median	7,7	4,9			Median	20,3	5,2	
	75%	13,7	6,5			75%	28,7	7,3	
Number of events	25%	4	9	< 0.001	Number of events	25%	4	5	< 0.001
	Median	6	20			Median	5	13	
	75%	9	38			75%	7	25	
Number of events per year	25%	0,5	2,9	< 0.001	Number of events per year	25%	0,2	1,7	< 0.001
	Median	0,8	5,4			Median	0,3	3,3	
	75%	1,6	10,0			75%	0,4	7,6	
Number of unique medications	25%	10	14	< 0.001	Number of unique diagnoses	25%	5	5	< 0.001
	Median	14	23			Median	7	9	
	75%	19	34			75%	9	15	
Interval between events (Days)	25%	113,3	3,0	< 0.001	Interval between events (Days)	25%	464,6	13,0	< 0.001
	Median	288,0	13,5			Median	914,0	31,0	
	75%	576,5	41,0			75%	1826,0	89,0	

Table 2. Comparison of longitudinality and sampling density variables between concomitant medication and diagnosis domains for the RCT ($N = 5\,734$) and RWD ($N = 23\,523$) patients. *Presented P values were adjusted with Bonferroni correction to address the issue of multiple comparisons.
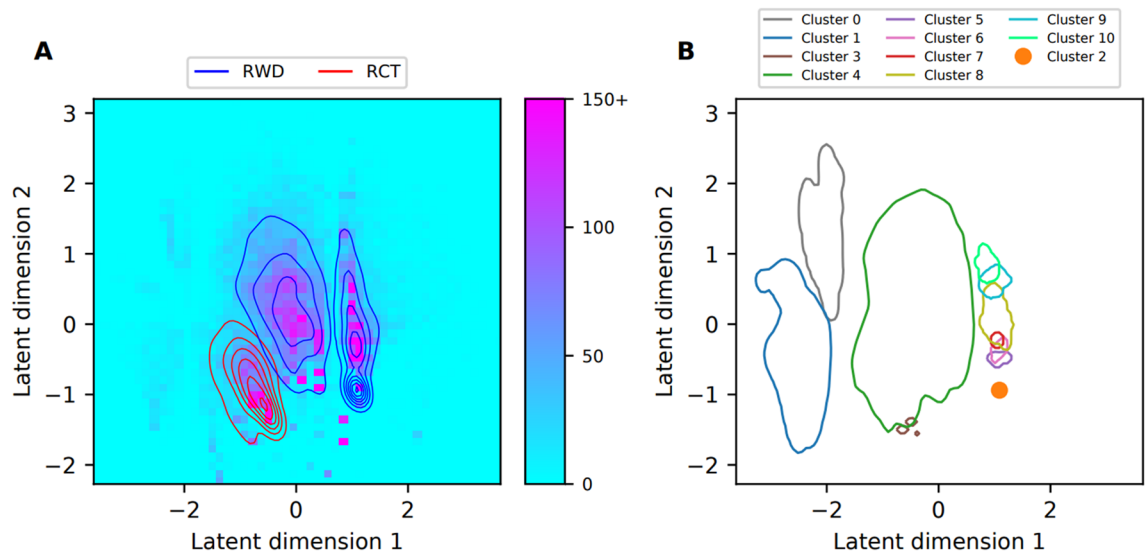

Figure 2. Visualization of the results from the cluster analysis of the diagnosis data. (**A**) The two-dimensional histogram shows the density of the combined RCT and RWD data sets along the two dimensions of the learned latent representation. We truncated histogram bin counts to 150. The red and blue lines show the Gaussian kernel density estimates (KDE) of the RCT and RWD distributions, respectively. (**B**) Contour lines for the Gaussian KDEs fitted using the datapoints belonging to clusters 0 and 3–10, plotted at the 0.05 value of each probability density estimate. The datapoints in cluster 2 lie very near to each other, leading to kernel density estimation failing, which is why we visualized the mean of these data point locations instead.

presence of both RCT and RWD patients within a single cluster. Importantly, over half of both the RCT and RWD cohorts were consolidated within Cluster 4, whereas the proportions for other clusters were substantially lower. Clusters 2 and 5 were distinguished by a very low prevalence for all diagnoses utilized in VAE model training.

Aside from these two clusters, those predominantly composed of RWD patients could be loosely divided into two groups: the larger Cluster 4 containing 17,383 patients, and a group comprising the adjacent Clusters 6–10. In Cluster 4, essential hypertension (I10) and non-insulin-dependent diabetes (E11) were the most prevalent diagnoses. Clusters 7–10 were characterized by prevalent diagnoses of pneumonia (J18) and either heart failure (I150), acute myocardial infarction (I21), or both. Cluster 6 differed substantially from Clusters 7–10, although geographically close. The most common diagnoses for Cluster 6 were other diseases of the urinary system (N39) and other soft tissue disorders (M79).
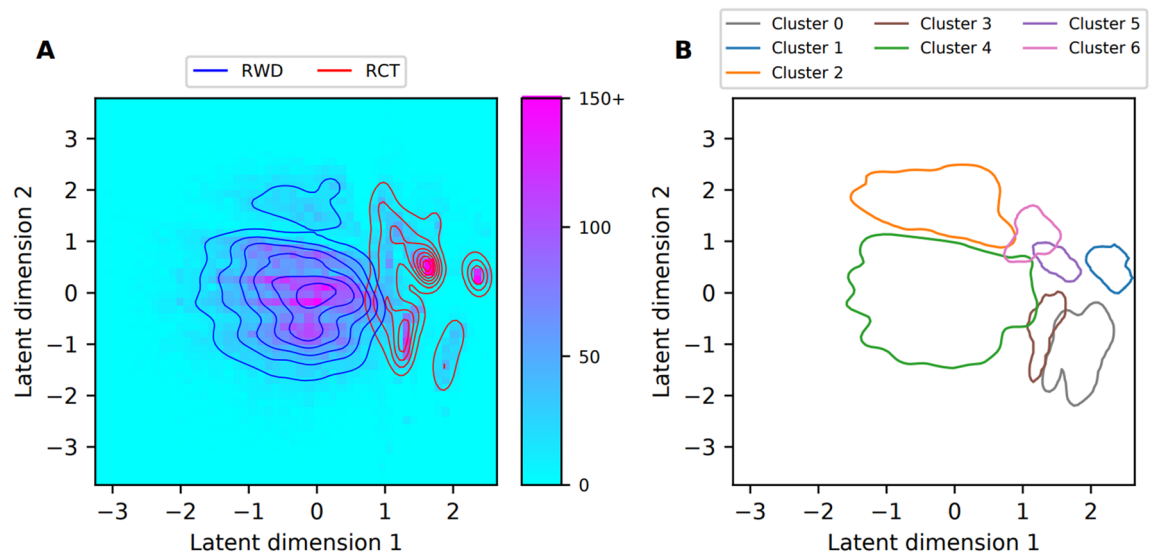
Figure 3. Visualization of the results from the cluster analysis of the medications data. (**A**) The two-dimensional histogram shows the density of the combined RCT and RWD data sets along the two dimensions of the learned latent representation. We truncated histogram bin counts to 150. The red and blue lines show the Gaussian kernel density estimates (KDE) of the RCT and RWD distributions, respectively. (**B**) Contour lines for the Gaussian KDEs fitted using the datapoints belonging to each cluster, plotted at the 0.05 value of each probability density estimate.

The RCT-dominated clusters (0, 1, and 3) shared prevalent diagnosis codes for chronic renal failure (N18), non-insulin-dependent diabetes (E11), and essential hypertension (I10). Additional prevalent diagnoses in Cluster 0, but not in Cluster 1, included vascular dementia (F01) and depressive episode (F32). Cluster 1, but not Cluster 0, exhibited a higher prevalence of other specified diabetes mellitus (E13), unspecified diabetes mellitus (E14), and glomerular disorders in diseases classified elsewhere (N08).

We further inspected the similarity of diagnosis prevalences within Clusters 3 and 4 using odds ratio and chi-squared test (Supplementary Material). In Cluster 3, the prevalences of chronic renal failure (N18) and non-insulin-dependent diabetes mellitus (E11) were in balance between RCT and RWD patients with odds ratio of 1.0 and $p = 1.0$, in contrast to respective odds ratios of 5.7 and 1.9 in the original datasets. The same trend towards more balanced prevalences was observed in the diagnoses of Cluster 4 compared to datasets without clustering.

From the medication dataset, two clusters (2 and 4) were mostly composed of RWD patients, and five (0, 1, 3, 5, and 6) were largely comprised of RCT patients. The most substantial overlap between RCT and RWD cohorts was observed in Cluster 0 (11% RWD, 89% RCT) and the largest Cluster 4 (5% RCT, 95% RWD). Over 70% of the RWD patients were included in Cluster 4, and the proportions for the remaining clusters were significantly smaller.

Conversely, RCT patients were more evenly distributed across the clusters Clusters 3, 4, and 5 each comprised approximately 15% of the RCT patients, with the remaining clusters containing smaller proportions. Across all clusters, the most frequently prescribed medications (prevalence > 60%) incorporated at least one medication code pertaining to the cardiovascular system (ATC codes beginning with 'C') and at least one code related to diabetes medications (ATC codes beginning with 'A*').

Contrary to other clusters, neither antithrombotic agents (B01A) nor opioids (N02A) and other analgesics and antipyretics (N02B) were prevalent in Clusters 0 and 3. Clusters 1, 5, and 6 all shared stomatological preparations (A01A) as their most frequent medication code, whereas this code was scarcely seen in Clusters 2 and 4. Clusters 2 and 4 were distinctive in that they had high prevalences of other beta-lactam antibacterials (J01D) and hypnotics and sedatives (N05C). In addition, other antianemic preparations (B03X) and calcium (A12A) were among the medications specific to Cluster 2.

Discussion

This study illustrates that real-world data (RWD) and randomized controlled trial (RCT) datasets, derived from patients with diabetic chronic kidney disease, share common characteristics but also exhibit substantial differences in terms of data generation, completeness, and temporal dynamics. These discrepancies have implications for study design validity and mandate careful examination when merging RWD and RCT data.

Data generation and longitudinality

The noted differences between RCT and RWD predominantly arise from their respective data generation processes and objectives. RCT data is prospectively collected following a specified study protocol, whereas RWD is extracted through queries from hospital data infrastructure. For instance, in RCT data, only the initial diagnosis date is recorded by the investigator on case report forms (CRFs), potentially leading to selection and recall bias.

Cluster	Cluster size		Proportion out of data set		Cluster breakdown		Most prevalent diagnosis codes**
	Number of patients	Proportion out of all patients*	RWD	RCT	RWD	RCT	
Patients without cluster assignment	5133	18.1%	20.5%	8.9%	90.2%	9.8%	–
0	366	1.3%	0.0%	6.4%	1.1%	98.9%	Chronic renal failure (N18), Vascular dementia (F01), Non-insulin-dependent diabetes mellitus (E11), Depressive episode (F32), Essential (primary) hypertension (I10), Obesity (E66), Disorders of lipo-protein metabolism and other lipidaemias (E78), Chronic ischaemic heart disease (I25), Sleep disorders (G47), Other anaemias (D64)
1	562	2.0%	0.0%	9.8%	1.4%	98.6%	Non-insulin-dependent diabetes mellitus) (E11), Chronic renal failure (N18), Unspecified diabetes mellitus (E14), Other specified diabetes mellitus (E13), Glomerular disorders in diseases classified elsewhere (N08), Essential (primary) hypertension (I10), Disorders of lipoprotein metabolism and other lipidaemias (E78), Obesity (E66), Chronic ischaemic heart disease (I25), Other cataract (H26)
2	1622	5.7%	7.2%	0.0%	100.0%	0.0%	No diagnoses with prevalence > 20%
3	453	1.6%	0.2%	7.1%	10.8%	89.2%	Chronic renal failure (N18), Non-insulin-dependent diabetes mellitus (E11), Essential (primary) hypertension (I10)
4	17,383	61.5%	59.9%	67.8%	78.0%	22.0%	Essential (primary) hypertension (I10), Non-insulin-dependent diabetes mellitus (E11), Chronic renal failure (N18), Disorders of lipoprotein metabolism and other lipidaemias (E78), Heart failure (I50), Chronic ischaemic heart disease (I25), Pneumonia, organism unspecified (J18)
5	271	0.1%	1.2%	0.0%	100.0%	0.0%	No diagnoses with prevalence > 20%
6	264	0.1%	1.2%	0.0%	100.0%	0.0%	Other diseases of urinary system (N39), Other soft tissue disorders (M79)
7	399	1.4%	1.8%	0.0%	100.0%	0.0%	Pneumonia, organism unspecified (J18), Acute myocardial infarction (I21)
8	1164	4.1%	5.1%	0.0%	100.0%	0.0%	Pneumonia, organism unspecified (J18), Heart failure (I50)
9	386	1.4%	1.7%	0.0%	100.0%	0.0%	Heart failure (I50), Pneumonia, organism unspecified (J18)
10	281	0.1%	1.2%	0.0%	100.0%	0.0%	Heart failure (I50), Pneumonia, organism unspecified (J18), Nonrheumatic aortic valve disorders (I35), Acute myocardial infarction (I21), Other chronic obstructive pulmonary disease (J44)

Table 3. Results from clustering of diagnosis data. *All patients used for the clustering of diagnosis data set. ** Diagnoses with at least 20% prevalence in the cluster are listed in descending order of prevalence out of 65 diagnosis codes used for VAE model training. Proportion out of data set describes how the RWD and RCT patients are distributed along the different clusters. Cluster breakdown shows how large proportions of the patients belonging to a cluster come from RWD and RCT sets.

Conversely, in RWD, each inpatient and outpatient diagnosis are precisely dated; however, data from a single institution may not encompass the patient's entire medical history. Unlike RWD, RCT data aims to assess the efficacy and safety of an intervention. Therefore, data elements unrelated to the trial's exposure and outcome, such as anti-infective medications and symptom diagnoses in our study, may be underrepresented. Furthermore, specific elements of RCT data, such as laboratory measurements, may only be cross-sectional and timed near the index date. Conversely, electronic health record (EHR) data chronicles patient interaction with healthcare services and inherently contains longitudinal data without defined start or end dates, possibly spanning decades. In our study, our research permit limited RWD to a ten-year range. Thus, incomplete patient history, whether derived from CRFs or EHRs, introduces biases and differences between RCT and RWD data. Ideally, RCT baseline data should utilize RWD covering the complete patient history over the relevant time range.

Data density and completeness

Diagnoses and medications were sampled significantly more densely in our RWD set compared to RCT. Consequently, RWD offered a more accurate portrayal of the patients' state by capturing all pertinent data. Nevertheless, data completeness and accuracy can be limited if data is sourced solely from a single healthcare provider. Despite all patients meeting the inclusion criteria, some lacked records of chronic kidney disease or type 2 diabetes diagnosis in RWD data, unlike in the RCT data. These diagnoses may be partly recorded in primary healthcare data,

Cluster	Cluster size		Proportion out of data set		Cluster breakdown		
	Number of patients	Proportion out of all patients*	RWD	RCT	RWD	RCT	Most prevalent medication codes**
Patients without cluster assignment	5487	19.4%	17.4%	27.5%	71.6%	28.4%	–
0	630	2.2%	0.3%	9.9%	11.0%	89.0%	Ace inhibitors, plain (C09A), Blood glucose lowering drugs, excl. insulins (A10B), Lipid modifying agents, plain (C10A)
1	453	1.6%	0.0%	8.0%	0.0%	100.0%	Stomatological preparations (A01A), Topical products for joint and muscular pain (M02A), Antithrombotic agents (B01A), ACE inhibitors, plain (C09A), Other analgesics and antipyretics (N02B), Lipid modifying agents, plain (C10A), Blood glucose lowering drugs, excl. insulins (A10B), Insulins and analogues (A10A), Beta blocking agents (C07A)
2	1674	5.9%	7.4%	0.2%	99.5%	0.5%	Other antianemic preparations (B03X), Other analgesics and antipyretics (N02B, Antithrombotic agents (B01A), Opioids (N02A, High-ceiling diuretics (C03C), Drugs for peptic ulcer and gastro-oesophageal reflux disease (GORD) (A02B), Calcium (A12A), Insulins and analogues (A10A), Selective calcium channel blockers with mainly vascular effects (C08C), Beta blocking agents (C07A), Drugs for constipation (A06A), Lipid modifying agents, plain (C10A), Other beta-lactam antibacterials (J01D), Vitamin A and D, incl. combinations of the two (A11C), Antiemetics and antinauseants (A04A), Corticosteroids for systemic use, plain (H02A), Iron preparations (B03A), Hypnotics and sedatives (N05C), Anxiolytics (N05B), Beta-lactam antibacterials, penicillins (J01C)
3	875	3.1%	0.1%	15.1%	2.6%	97.4%	Angiotensin II receptor blockers (ARBs), plain (C09C), Blood glucose lowering drugs, excl. insulins (A10B), Lipid modifying agents, plain (C10A), Selective calcium channel blockers with mainly vascular effects (C08C), Insulins and analogues (A10A)
4	17,751	62.9%	74.8%	15.1%	95.2%	4.8%	Antithrombotic agents (B01A), Other analgesics and antipyretics (N02B), Beta blocking agents (C07A), Opioids (N02A), Lipid modifying agents, plain (C10A), High-ceiling diuretics (C03C), Drugs for peptic ulcer and gastro-oesophageal reflux disease (GORD) (A02B), Other beta-lactam antibacterials (J01D), Insulins and analogues (A10A), Blood glucose lowering drugs, excl. insulins (A10B), Hypnotics and sedatives (N05C), Drugs for constipation (A06A), Selective calcium channel blockers with mainly vascular effects (C08C)
5	937	3.3%	0.0%	16.6%	0.0%	100.0%	Stomatological preparations (A01A), Antithrombotic agents (B01A), Other analgesics and antipyretics (N02B), Topical products for joint and muscular pain (M02A), Angiotensin II receptor blockers (ARBs), plain (C09C), Lipid modifying agents, plain (C10A), Blood glucose lowering drugs, excl. insulins (A10B), Selective calcium channel blockers with mainly vascular effects (C08C), Insulins and analogues (A10A)
6	436	1.5%	0.0%	7.7%	0.7%	99.3%	Stomatological preparations (A01A), Lipid modifying agents, plain (C10A), Antithrombotic agents (B01A), Topical products for joint and muscular pain (M02A), Other analgesics and antipyretics (N02B), Blood glucose lowering drugs, excl. insulins (A10B), Selective calcium channel blockers with mainly vascular effects (C08C), Insulins and analogues (A10A), Angiotensin II receptor blockers (ARBs), plain (C09C)

Table 4. Results from clustering of medications data. *All patients used for the clustering of medication data set. **Medications with at least 60% prevalence in the cluster are presented in descending order of prevalence out of 84 medication codes used for VAE model training. Proportion out of data set describes how the RWD and RCT patients are distributed along the different clusters. Cluster breakdown shows how large proportions of the patients belonging to a cluster come from RWD and RCT sets.

which was not included in this study. Additionally, text mining was necessary to extract data from unstructured texts in RWD. Despite these limitations, our findings suggest that, in certain scenarios, RWD could supplement RCT data through EHR to electronic data capture (EHR2EDC) automation[30].

The harmonization process resulted in a minor loss of diagnoses, with 94.1% of RWD and 95.9% of RCT diagnosis codes successfully mapped to SNOMED codes. The mapping process for the remaining data types was straightforward.

Cluster analysis

We utilized cluster analysis on the combined real-world and randomized controlled trial (RCT) datasets to illustrate the heterogeneity of the study population, discover patient subgroups, and assess the overlap between the two datasets. The clustering was essentially based on the idea that Variational Autoencoders can learn nonlinear mapping between high-dimensional feature vectors representing patient characteristics and low-dimensional latent space representation. Similar feature vectors are basically located close to each other in the latent space. Our analysis revealed that the datasets were largely distinct. Both RCT and real-world data (RWD) sets comprised unique subgroups, with an overlap observed in only a few clusters. This was true for both diagnosis and medication datasets. However, clustering enabled us to extract from RCT and RWD datasets subgroups of patients who had more similar characteristics than original datasets. In the diagnosis data, Cluster 3 and the largest cluster (Cluster 4) contained a significant number of patients from both RWD and RCT sets. Hence, even though overlap was not present in all clusters, many RWD and RCT patients were grouped together in the cluster analysis. Clusters represented different patient characteristics and tended towards balanced prevalences of features between RCT and RWD sets in cluster-specific manner. Thus, clustering was found useful in understanding and mitigating group differences by selecting subgroups of patients with similar characteristics. Due to low availability of specific parameters in RWD, we did not apply the whole set of trial criteria, which is possibly reflected in the extent of overlap between the datasets. On the other hand, RWD contained a wider spectrum of recorded diagnoses than RCT. We conclude that much of the observed differences are due to different data generation mechanisms of RWD and RCT data.

The cluster analyses underscore the challenges in finding overlaps between real-world and RCT data, emphasizing the necessity for advanced methods to identify matching external controls. In the cluster analyses, we chose input covariates using a prevalence threshold of 1%, leading to 65 and 84 input covariates for diagnoses and medications data, respectively. A higher threshold would yield fewer covariates that could potentially differentiate the real-world and RCT datasets. The covariate set chosen for aligning RCT and RWD can influence the outcome and therefore requires careful selection. In addition to clinical differences between groups, there could be disparities in data completeness, i.e., how well the selected covariates are captured in the different data sources. It's important to note that the clustering results are influenced by our choices made in the VAE model training and cluster analysis and represent one set of possible options. Furthermore, we conducted the cluster analyses without considering possible demographic differences between RCT and RWD, which could account for some of the identified differences. To assess the robustness of clustering, subsampling-based analysis as in[22] would be preferable. However, cluster analysis proved beneficial in identifying patients with overlapping characteristics in RCT and RWD. Our results indicate the overlap and discrepancies in one trial and RWD pair but cannot be directly generalized. Nevertheless, similar observations are likely in any study that merges RCT and RWD sources. The clustering method could be used to identify the characteristics of the most common phenotype of patients in a selected therapeutic area of a healthcare provider and compare it to the characteristics of the patients in a global clinical trial. On the other hand, the method could be used to identify more rare patient phenotypes to enable application of precision medicine and further expansion of the research to rare subpopulations of a disease.

Conclusion

In this study, we successfully elucidated the differences and demonstrated the feasibility of combining RCT and RWD, highlighting the potential for enriching RCT data using first-hand baseline information, filling missing data, and effectively mitigating discrepancies between datasets. RCT and RWD exhibit substantial differences in data longitudinality, completeness, sampling density, among other factors, all of which should be considered when designing studies that amalgamate data from these sources. Despite their inherent limitations, RWD sources could be used to enrich RCT datasets, for instance, to enhance the longitudinality and completeness of patient history. RCT and RWD sets were distinct and could form unique patient subgroups, which must be considered in studies merging RCT and RWD and in patient matching.

Data availability

Regarding the RWD, according to the Finnish legislation, access to individual-level data is restricted only to individuals named in the study permit. The study protocol is available upon request from the corresponding author. Regarding the RCT data, the data are not publicly available due to containing information that could compromise research participant privacy/consent.

References
 1. Abrahami, D. *et al.* Use of real-world data to emulate a clinical trial and support regulatory decision making: Assessing the impact of temporality, comparator choice, and method of adjustment. *Clin. Pharmacol. Ther.* **109**(2), 452–461 (2021).

2. Jemielita, T. *et al.* Overall survival with second-line pembrolizumab in patients with non–small-cell lung cancer: Randomized phase III clinical trial versus propensity-adjusted real-world data. *JCO Clin. Cancer Inform.* **5**, 56–65 (2021).
3. Chen, Z. *et al.* Exploring the feasibility of using real-world data from a large clinical data research network to simulate clinical trials of Alzheimer's disease. *NPJ Digit. Med.* **4**(1), 1–9 (2021).
4. Rogers, J. R. *et al.* Contemporary use of real-world data for clinical trial conduct in the United States: A scoping review. *J. Am. Med. Inform. Assoc.* **28**(1), 144–154 (2021).
5. Carrigan, G. *et al.* Using electronic health records to derive control arms for early phase single-arm lung cancer trials: Proof-of-concept in randomized controlled trials. *Clin. Pharmacol. Ther.* **107**(2), 369–377 (2020).
6. Davies, J. *et al.* Comparative effectiveness from a single-arm trial and real-world data: Alectinib versus Ceritinib. *J. Comp. Eff. Res.* **7**(9), 855–865 (2018).
7. Jemielita, T. *et al.* Replication of oncology randomized trial results using Swedish registry real world-data: A feasibility study. *Clin. Pharmacol. Ther.* **110**(6), 1613–1621 (2021).
8. Patry, C. *et al.* Emulation of the control cohort of a randomized controlled trial in pediatric kidney transplantation with Real-World Data from the CERTAIN Registry. *Pediatr. Nephrol.* **38**(5), 1621–1632 (2023).
9. Popat, S. *et al.* Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of Pralsetinib in non-small cell lung cancer. *Nat. Commun.* **13**(1), 1–10 (2022).
10. Lin, K. J. *et al.* Out-of-system care and recording of patient characteristics critical for comparative effectiveness research. *Epidemiology* **29**(3), 356–363 (2018).
11. Franklin, J. M. & Schneeweiss, S. When and how can real world data analyses substitute for randomized controlled trials?. *Clin. Pharmacol. Ther.* **102**(6), 924–933 (2017).
12. Rassen, J. A., Bartels, D. B., Schneeweiss, S., Patrick, A. R. & Murk, W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clin. Epidemiol.* **17**, 1–5 (2018).
13. Suissa, S. & Dell'Aniello, S. Time-related biases in pharmacoepidemiology. *Pharmacoepidemiol. Drug Safety* **29**(9), 1101–1110 (2020).
14. Ghadessi, M. *et al.* A roadmap to using historical controls in clinical trials–by drug information association adaptive design scientific working group (DIA-ADSWG). *Orphanet J. Rare Dis.* **15**(1), 1–19 (2020).
15. Nikram, B. & Zubizarreta, J. Using cardinality matching to design balanced and representative samples for observational studies. *JAMA* **327**(2), 173–174 (2022).
16. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55 (1983).
17. Bakris, G. *et al.* Effect of Finerenone on chronic kidney disease outcomes in type 2 diabetes. *New Eng. J. Med.* **383**, 2219–2229 (2020).
18. Athena, Athena.ohdsi.org, [Online]. Available: https://athena.ohdsi.org/search-terms/start. [Accessed 17 April 2023]. (2023)
19. E. Kuorttinen and M. Salmikangas, HUS Acamedic. Virtual presentation, [Online]. Available: https://sway.office.com/iiBUeFL0ZCF3ChY2?ref=Link. [Accessed 27 March 2023]. (2022)
20. W. McKinney, Data Structures for Statistical Computing in Python, *Proceedings of the 9th Python in Science Conference,* pp. 56–61, (2010).
21. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods,* no. 17, pp. 261–272, (2020).
22. Koskinen, M. *et al.* Data-driven comorbidity analysis of 100 common disorders reveals patient subgroups with differing mortality risks and laboratory correlates. *Sci. Rep.* **12**(1), 1–9 (2022).
23. D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes, *International Conference on Learning Representations,* (2014).
24. D. J. Rezende, S. Mohamed and D. Wierstra, Stochastic Backpropagation and Approximate Inference in Deep Generative Models, *Proceedings of the 31st International Conference on Machine Learning,* vol. 32, no. 2, pp. 1278–1286, (2014).
25. F. Chollet and and others, Keras, [Online]. Available: https://keras.io. [Accessed 27 March 2023]. (2015)
26. L. McInnes, J. Healy and S. Astels, hdbscan: Hierarchical density based clustering, *Journal of Open Source Software,* vol. 2, no. 11, (2017).
27. I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, (2016).
28. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
29. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**(3), 90–95 (2007).
30. Cowie, M. R. *et al.* Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **106**, 1–9 (2017).

Author contributions

Conception of the work: MK, JL Study design: SK, VH, JL, MK Acquisition of data: SK, VH, JL, MK Analysis: SK, VH Interpretation: SK, VH, MH, HL, JL, MK Drafting the manuscript text: SK, VH, JL, MK Revision of the manuscript critically for important intellectual content: SK, VH, MH, HL, JL, MK Final approval of the version to be published: SK, VH, MH, HL, JL, MK Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: SK, VH, MH, HL, JL, MK.

Competing interests

Samu Kurki and Jussi Leinonen are employed by Bayer Oy (Finland). All other authors declare that they do not have any competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-51938-3.

Correspondence and requests for materials should be addressed to S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.