1. Title Page

Title	The effect of SGLT2 inhibitors among patients with Type 2 diabetes on the incidence of colorectal cancer
Research question & Objectives	To estimate the effect of incident use of SGLT2 inhibitors on the risk of newly diagnosed colorectal cancer compared with incident use of DPP4 inhibitors.
Protocol version	1.0
Last update date	27 January 2026
Contributors	**Primary investigator contact information:** Britta Haenisch Britta.haenisch@bfarm.de **Contributor names:** Martin Russek Julia Wicherski Andreas Brandt Ulrike Hermes Frauke Naumann-Winter
Study registration	**Site:** HMA-EMA catalogue of RWD studies **Identifier:** 1000000920
Sponsor	Funding from the European Community's Horizon Europe Programme under grant agreement No. 101095353 (Real4Reg)
Conflict of interest	None

Table of contents

2. Abstract

The diabetes medication class of SGLT2 inhibitors (SGLT2i) has been shown to not only improve glycaemic control in type 2 diabetes patients, but also provide cardiovascular benefit in both diabetic patients[1] and patients with heart failure,[2] as well as renal benefit.[3] Recent nonclinical research results have additionally suggested a potential benefit in reducing colorectal tumour growth.[4–6] A retrospective cohort study from Hong Kong has supported this claim by estimating a reduced risk of incident colorectal cancer (CRC) for SGLT2i users compared to DPP4 inhibitor (DPP4i) users.[7] Two Asian multi-outcome database studies suggest a similar protective effect.[8,9]

This study aims to estimate the causal effect of initiating treatment with SGLT2i versus DPP4i additionally to metformin therapy in persons with type 2 diabetes. It is designed as a cohort study based on German health insurance claims data from the years 2009 – 2023; targeted maximum likelihood estimation for survival & competing risk analysis will be used to estimate absolute CRC-specific risk curves for both treatment groups and 5-year average treatment effects will be estimated.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason

4. Milestones

Table 1 Milestones

Milestone	Date
Data access	April 2026
Analysis completed	June 2026
Manuscript finalised	October 2026

5. Rationale and background

What is known about the condition: Colorectal cancer is the second most diagnosed cancer for women and the third most diagnosed for men,[10] and it is responsible for almost 10% of all cancer-related mortality worldwide.[11] There are several factors contributing to the development of colorectal cancer, including uncontrolled blood glucose levels.[12] Thus, it has been speculated that certain diabetes medications might also have a protective effect against colorectal cancer.

What is known about the exposure of interest: SGLT2i act by inhibiting the protein SGLT2, which is involved in glucose reabsorption. Through this activity, blood glucose is reduced.

Gaps in knowledge: The effect of SGLT2i on development of colorectal cancer

What is the expected contribution of this study? Exploratory evidence on potential protective effects of SGLT2i against the development of colorectal cancer.

6. Research question and objectives

Table 2 Primary and secondary research questions and objective

A. **Primary research question and objective**

Objective:	To estimate the effect of incident use of SGLT2 inhibitors on the risk of newly diagnosed colorectal cancer compared with incident use of DPP4 inhibitors
Hypothesis:	SGLT2i use reduces the risk of incident colorectal cancer diagnoses
Population *(mention key inclusion-exclusion criteria):*	Persons with Type 2 diabetes already using metformin, at least 30 years of age, who have never been diagnosed with cancer (excl. non-melanoma skin cancer) and have never been treated with SGLT2i or DPP4i before study entry
Exposure:	Incident use of a SGLT2i
Comparator:	Incident use of a DPP4i
Outcome:	Time to incident diagnosis of colorectal cancer
Time *(when follow up begins and ends):*	Treatment initiation until earliest of outcome, death, other cancer diagnosis or loss to follow-up
Setting:	Inpatient (diagnosis), outpatient (treatment & diagnosis)

Main measure of effect:	5-year risk difference

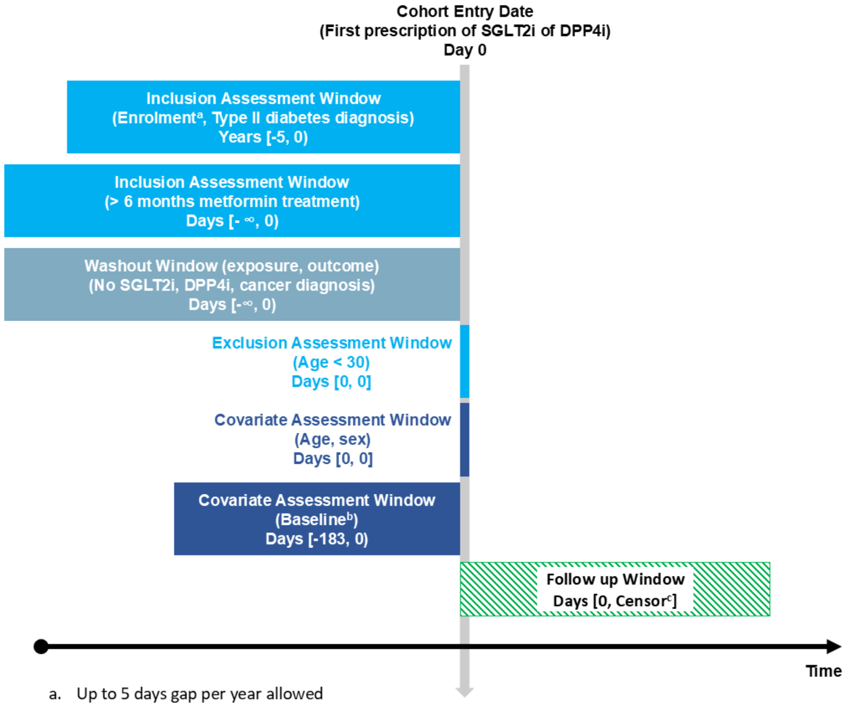
7. Research methods

7.1. Study design

Research design (e.g. cohort, case-control, etc.): Observational cohort study; Active comparator new user (ACNU) design

Rationale for study design choice: Best suited for establishing potential causal claims in observational studies; randomised controlled trial infeasible. Sample size relatively large and outcome sufficiently common, thus case-control study not indicated.

7.2. Study design diagram


Cohort Entry Date
(First prescription of SGLT2i of DPP4i)
Day 0
Inclusion Assessment Window
(Enrolment[a], Type II diabetes diagnosis)
Years [-5, 0)
Inclusion Assessment Window
(> 6 months metformin treatment)
Days [- ∞, 0)
Washout Window (exposure, outcome)
(No SGLT2i, DPP4i, cancer diagnosis)
Days [-∞, 0)
Exclusion Assessment Window
(Age < 30)
Days [0, 0]
Covariate Assessment Window
(Age, sex)
Days [0, 0]
Covariate Assessment Window
(Baseline[b])
Days [-183, 0)
Follow up Window
Days [0, Censor[c]]
Time

a. Up to 5 days gap per year allowed
b. Baseline covariates include: Demographics, lifestyle, health & diagnoses, procedures and medications. See appendix for detailed list
c. Earliest of: outcome of interest (CRC), other cancer diagnosis, death, disenrollment, end of the study period

7.3. Setting

7.3.1 Context and rationale for definition of time 0 (and other primary time anchors) for entry to the study population

Time 0 for both exposed and control group are defined as the date of incident prescription of SGLT2i (ATC A10BK) or DP44i (ATC A10BH). Users have to be incident with respect to both medications. This time 0 was selected to correspond to the target trial exposure groups of being assigned to start treatment with SGLT2i or DPP4i as an additional treatment line for people that have not received any of the two treatments previously.

Table 3 Operational Definition of Time 0 (index date) and other primary time anchors

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm

Exposed	Date of incident prescription of a SGLT2i	Single entry	Incident	(-∞, 0)	OP	ATC		SGLT2i and DPP4i use		
Comparator	Date of incident prescription of a DPP4i	Single entry	Incident	(-∞, 0)	OP	ATC		SGLT2i and DPP4i use		

[1] IP = inpatient, OP = outpatient

[2]See appendix for listing of clinical codes for each study parameter

7.3.2 Context and rationale for study inclusion criteria:

Individuals need to have continuous enrolment within the data source during lookback, to ensure all relevant healthcare contacts are captured. Only patients at least 30 years of age at time of diagnosis, to minimize the risk of false inclusion of Type 1 Diabetes patients. Diagnosis of Type 2 Diabetes and treatment with metformin are used to define the population of interest.

Table 4. Operational Definitions of Inclusion Criteria

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Enrolment	Continuous enrolment in data source during lookback period. Gap of 5 days per year allowed.	After	[-5y, 0)	NA	NA	NA	All		
Age	Age at index date at least 30. Age in years defined by (time 0 – year of birth)	After	0	NA			All		
Type II diabetes	Diagnosis with Type 2 diabetes (E11). One diagnosis suffices, due to additional inclusion criterion 'Metformin treatment'	After	[-5y, 0)	IP, OP	ICD-10-GM	Outpatient: confirmed (G); Inpatient: any	All		
Metformin treatment	Prescription of metformin in the 6 months before index date	After	[-183d, 0)	OP	ATC		All		

[1] IP = inpatient, OP = outpatient, NA = not applicable
[2] See appendix for listing of clinical codes for each study parameter

7.3.3 Context and rationale for study exclusion criteria

Individuals with previous SGLT2i/DPP4i treatment are excluded, as only incident treatment with the two drug classes is of interest in this study. Similarly for previous cancer diagnoses.

Table 5. Operational Definitions of Exclusion Criteria

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Previous treatment with DPP4i or SGLT2i	Prescription of DPP4i or SGLT2i at any time before index date	After	(-∞, 0)	OP	ATC		All		
Previous cancer diagnosis	Diagnosis of cancer (excl. non-melanoma skin cancer) at any time prior to index date. At least 2 confirmed diagnosis codes in outpatient setting or at least one primary or secondary diagnosis in inpatient setting.	After	(-∞, 0)	IP, OP	ICD-10-GM	Outpatient: confirmed (G); Inpatient: any	All		

[1] IP = inpatient, OP = outpatient

[2] See appendix for listing of clinical codes for each study parameter

7.4. Variables

7.4.1 Context and rationale for exposure(s) of interest

As SGLT2i were the exposure of interest, DPP4i were selected as comparators, since both drugs are given as 2nd+-line therapy for type 2 diabetes and both are given orally. Both drug exposures are defined as single-timepoint exposures – the comparison of interest is in the effect of choosing SGLT2i rather than DPP4i as 2nd+-line treatment.

Table 6. Operational Definitions of Exposure

Exposure group name(s)	Details	Washout window	Assessment Window	Care Setting[1]	Code Type[2]	Diagnosis position	Applied to study populations:	Incident with respect to...	Measurement characteristics/ validation	Source of algorithm
SGLT2i	Prescription of any SGLT2i in any formulation and amount	(-∞, 0)	0	OP	ATC		Exposure	SGLT2i and DPP4i		
DPP4i	Prescription of any DPP4i in any formulation and amount	(-∞, 0)	0	OP	ATC		Exposure	SGLT2i and DPP4i		

[1] IP = inpatient, OP = outpatient

[2] See appendix for listing of clinical codes for each study parameter

7.4.2 Context and rationale for outcome(s) of interest

The outcome of interest is incident diagnosis of CRC.

Table 7. Operational Definitions of Outcome

Outcome name	Details	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Type[2]	Diagnosis Position	Applied to study populations:	Measurement characteristics/ validation	Source of algorithm
Colorectal cancer	Time to incident diagnosis of colorectal cancer (ICD10-GM C18-C20). Diagnoses within 6 months of CED censored as non-event	Yes	Time-to-event	(-∞, 0)	IP, OP	ICD10-GM	Outpatient: confirmed (G); Inpatient: main and secondary diagnoses (1/H, 2/N)	All		Pottegård 2016[13]

[1] IP = inpatient, OP = outpatient

[2] See appendix for listing of clinical codes for each study parameter

7.4.3 Context and rationale for follow up

Follow-up starts at the date of incident prescription of a SGLT2i or DPP4i, however any events within the first 6 months will be censored as non-events to minimize risk for protopathic bias.[14] Death and diagnosis with other primary cancers will result in end of follow-up and will be treated as competing risks. Individuals not experiencing an event will be censored at the end of their observation in the data set or end of the study period. Patients will be followed irrespectively of treatment discontinuation or switch.

Table 8. Operational Definitions of Follow Up

Follow up start	First prescription of SGLT2i/DPP4i	
Follow up end[1]	**Select all that apply**	**Specify**
Date of outcome	Yes	
Date of death	Yes	Competing risk
End of observation in data	Yes	

Day X following index date *(specify day)*	No	
End of study period (specify date)	Yes	31 December 2023
End of exposure *(specify operational details, e.g. stockpiling algorithm, grace period)*	No	
Date of add to/switch from exposure *(specify algorithm)*	No	
Other date *(specify)*	Yes	Diagnosis with other cancer (competing risk); CRC diagnosis within 6 months of cohort entry (non-event)

[1] Follow up ends at the first occurrence of any of the selected criteria that end follow up.

7.4.4 Context and rationale for covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications)

Table 9. Operational Definitions of Covariates

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Sex	Sex at CED	Ternary (male, female, other)	0				All		
Age	Age in years at CED	Numeric	0				All		
Socioeconomic status	Z59 “Problems related to housing and economic circumstances”	Binary	[-356d, 0)	IP, OP	ICD-10-GM	Any	All		
CRC family history	Z80.0	Binary	(-∞, 0)	IP, OP	ICD-10-GM	Any	All		
Smoking	Nicotine dependence, prescription of drugs used in nicotine dependence	Binary	(-∞, 0)	IP, OP	ICD-10-GM, ATC	Outpatient: confirmed (G); Inpatient: any	All		
Supplemental calcium & vitamin D	Prescription of supplements	Binary	[-356d, 0)	OP	ATC		All		

Characteristic	Details	Type of variable	Assessment window	Care Settings¹	Code Type²	Diagnosis Position³	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Alcohol intake	Alcohol related disorders, prescription of drugs used in alcohol dependence	Binary	(-∞, 0)	IP, OP	ICD-10-GM, ATC	Outpatient: confirmed (G); Inpatient: any	All		
Diseases, Symptoms (15 individual Covariates)	Diagnoses	Binary	[-356d, 0)	IP, OP	ICD-10-GM	Outpatient: confirmed (G); Inpatient: any	All		
Medications (43 individual covariates)	Reimbursed medications	Binary	(-356d, 0)	OP	ATC		All		
Overall health (Travel fitness)	Use of travel vaccines	Binary	[-356d, 0)	OP	ATC		All		Herweijer 2022[15]
Cholecystectomy	Recorded procedure	Binary	(-∞, 0)	IP, OP	OPS		All		
Appendectomy	Recorded procedure	Binary	(-∞, 0)	IP, OP	OPS		All		
Medications	Prescriptions	Binary	[-356d, 0)	OP	ATC		All		
Healthcare utilisation	Number of healthcare contacts	numeric	[-356d, 0)	IP, OP			All		
Colonoscopy	Recorded procedure	Binary	[-356d, 0)	IP, OP	OPS		All		
Imaging of abdomen	Recorded procedure	Binary	[-356d, 0)	IP, OP	OPS		All		
Number of glucose lowering medicines	Number of prescriptions of distinct A10B classes	Numeric	[-356d, 0)	OP			All		
Frailty	R54	Binary	[-356d, 0)	IP, OP	ICD-10-GM	Any	All		
eGFR	GFR category in diagnosis	Categorical	[-356d, 0)	IP, OP	ICD-10-GM	Outpatient: confirmed (G); Inpatient: any	All		
Serum creatinine	Renal failure stadium in diagnosis	Categorical	[-356d, 0)	IP, OP	ICD-10-GM	Outpatient: confirmed (G); Inpatient: any	All		
Time since type II diabetes diagnosis	Diagnosis date - CED >= 5 years	Binary	(-∞, 0)	IP, OP			All		

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Year of prescription		Numeric	0	OP			All		

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.5. Data analysis

7.5.1 Context and rationale for analysis plan

Targeted maximum likelihood estimation (TMLE) in the context of competing risks[16] for time-to event data will be applied to estimate the CRC-specific absolute risk across time for the counterfactual scenarios ‘everybody is treated with SGLT2i’ and ‘everybody is treated with DPP4i’. Using these estimates, the average treatment effect of SGLT2i versus DPP4i initiation will be estimated 5 years after treatment initiation. TMLE was chose for the analysis as it is a doubly robust method, reducing the risk of bias from model misspecification. The average treatment effect is estimated as relative risk and risk difference between the two treatment groups – in contrast to the commonly used hazard ratios, the relative risk allows for a causal interpretation of the estimate. In addition to the 5-year ATE, the estimated survival curves will be compared globally, to identify changes to the ATE over time (including past 5 years)

Within TMLE, the super learner is used for predicting treatment probabilities. The algorithms included in the super learner are LASSO, random forests, gradient boosting and support vector machines. Those algorithms were selected to provide good model flexibility while accommodating the analysis platform’s infrastructure. Outcome and censoring modelling is done using the highly adaptive lasso (HAL) super learner, as this approach provides large flexibility towards distribution modelling.

Table 10. Primary, secondary, and subgroup analysis specification

A. **Primary analysis**

Hypothesis:	Initiating treatment with SGLT2i rather than DPP4i in patients with Type-2-Diabetes that have been treated with metformin leads to a reduced risk for incident colorectal cancer
Exposure contrast:	Incident treatment with SGLT2i vs. incident treatment with DPP4i
Outcome:	Incident CRC diagnosis
Analytic software:	R
Model(s): ***(provide details or code)***	TMLE 5-year average treatment effect of exposure defined above.

Confounding adjustment method	*Name method and provide relevant details, e.g. bivariate, multivariable, propensity score matching (specify matching algorithm ratio and caliper), propensity score weighting (specify weight formula, trimming, truncation), propensity score stratification (specify strata definition), other.*
	Targeted minimum loss-based estimation (TMLE). Covariates defined as above, with exposure distribution estimated via super learner and outcome and censoring distribution estimated via the highly adapted lasso (HAL) super learner
Missing data methods	*Name method and provide relevant details, e.g. missing indicators, complete case, last value carried forward, multiple imputation (specify model/variables), other.*
	Age and sex only variables for which concept of missing values exists, but both variables are mandatorily filled in the database. As data on procedures only exist from the years 2019, covariate information for the 4 variables involving OPS codes is missing until 2018. As it is not expected that the distribution of the 4 variables differs between treatment groups, those covariates will be dropped from the models; multiple imputation will only be considered if data from 2019 onwards suggests differences between treatment groups.

Table 11. Sensitivity analyses – rationale, strengths and limitations

What is being varied? How?	Why? (What do you expect to learn?)	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
Gap between treatment initiation and outcome assessment increased to 1 year and decreased to 3 months.	Impact of choice of gap period	Reduced risk for protopathic bias (1 year) or immortal time bias (3 months)	Decreased detection of short-term effects on CRC diagnosis (1 year); increased risk for protopathic bias (3 months)
GLP1-RA instead of DPP4i as comparator group	Impact of comparator choice	GLP1-RAs, like SGLT2i, may be given preferentially to patients with cardiovascular conditions	GLP1-RAs mostly administered as injection, opposed to SGLT2i and DPP4i being administered orally
While-on-treament approach to treatment switching and discontinuation	Impact of intercurrent event handling	Effect estimate closer to effect of treatment rather than the effect of choice of treatment	Long-term effects possibly underestimated
IPTW Fine-gray model instead of TMLE	Impact of model choice	Effect estimate (hazard ratio) more commonly used for survival analysis	Method less efficient than TMLE

7.6. Data sources

7.6.1 Context and rationale for data sources

Reason for selection: Drug exposure well-captured as both drugs are primarily prescribed in outpatient setting. Outcome well-captured. Largest claims data source in Europe, thus sufficient power to investigate study question

Strengths of data source(s): Large sample size

Limitations of data source(s): No inpatient prescriptions, no over-the-counter drug use captured. No cause-of-death data and granularity of cancer diagnoses limited to ICD-10.

Data source provenance/curation: GKV-SV/Forschungsdatenzentrum Gesundheit

Table 12. Metadata about data sources and software

	Data 1
Data Source(s):	Forschungsdatenzentrum Gesundheit (FDZ)
Study Period:	2009 - 2023
Eligible Cohort Entry Period:	2009 - 2023
Data Version (or date of last update):	TBD
Data sampling/extraction criteria:	Entire population
Type(s) of data:	Claims data
Data linkage:	Unique identifier
Conversion to CDM*:	No
Software for data management:	SQL, R

*CDM = Common Data Model

7.7. Data management

All data storage (including version control) and analysis is performed on the data sources own platform; no data transfer takes place.

7.8. *Quality control*

Data quality assurance is carried out by the data provider.

7.9. *Study size and feasibility*

In Germany in 2008, 68% of people with newly diagnosed T2D started with metformin, and 30% of those used DPP4i as 2nd line.[17] In 2021 there were more than 500.000 new T2D diagnoses per year in Germany. Thus, we expect around 100.000 people per year to be included in the study, assuing that SGLT2i users would have been starting on DPP4i if SGLT2i hadn't been available. According to Globocan 2022[10], the incidence of CRC at age 40-69 is 61/100.000 in Germany. Focusing on the 5-year relative risk, assuming a sample of 500.000 individuals, a baseline incidence of 61/100.000 and a 10% risk reduction, a 0.05 alpha-level would imply a power of 0.81.

8. Limitation of the methods

Since this in an observational study, no final conclusions on causality are possible, as residual confounding cannot be excluded despite use of modern causal inference methods. The data source used does not include data on lifestyle factors and overall physical fitness. However, it is unlikely that any lifestyle factors are directly related to treatment assignment, and indirect effects are captured via medical history. The data source does also not include data on prescriptions made in hospitals or over-the-counter medications. This has consequences on capturing some covariates adequately, but has no relevant impact on definition of treatment or outcomes. Diagnosis dates are only given by quarter or month, depending on year of record. As this study is investigating a long-term outcome, this also have no substantial impact on the interpretation of results.

In case there is substantial under-recording of relevant variables, or an unexpectedly large impact of unmeasured confounders, no unbiased estimation would be possible, regardless of the analysis method, and no valid scientific conclusion could be drawn from the results.

9. Protection of human subjects

Not applicable

10. Reporting of adverse events

Not applicable

11. References

1. Usman MS, Siddiqi TJ, Memon MM, et al. Sodium-glucose co-transporter 2 inhibitors and cardiovascular outcomes: A systematic review and meta-analysis. *Eur J Prev Cardiol*. 2018;25(5):495-502. doi:10.1177/2047487318755531

2. Anker SD, Butler J, Filippatos G, et al. Empagliflozin in Heart Failure with a Preserved Ejection Fraction. *N Engl J Med*. 2021;385(16):1451-1461. doi:10.1056/NEJMoa2107038

3. Suzuki Y, Kaneko H, Okada A, et al. Kidney outcomes with SGLT2 inhibitor vs. DPP4 inhibitor use in older adults with diabetes. *Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc*. Published online July 11, 2024:gfae158. doi:10.1093/ndt/gfae158

4. Saito T, Okada S, Yamada E, et al. Effect of dapagliflozin on colon cancer cell [Rapid Communication]. *Endocr J*. 2015;62(12):1133-1137. doi:10.1507/endocrj.EJ15-0396

5. Anastasio C, Donisi I, Del Vecchio V, et al. SGLT2 inhibitor promotes mitochondrial dysfunction and ER-phagy in colorectal cancer cells. *Cell Mol Biol Lett*. 2024;29:80. doi:10.1186/s11658-024-00599-1

6. Nasiri AR, Rodrigues MR, Li Z, Leitner BP, Perry RJ. SGLT2 inhibition slows tumor growth in mice by reversing hyperinsulinemia. *Cancer Metab*. 2019;7:10. doi:10.1186/s40170-019-0203-1

7. Chan RNC, Chan RNF, Chou OHI, Tse G, Lee S. Lower risks of incident colorectal cancer in SGLT2i users compared to DPP4i users: A propensity score-matched study with competing risk analysis. *Eur J Intern Med*. 2023;110:125-127. doi:10.1016/j.ejim.2023.01.021

8. Huang YM, Chen WM, Jao AT, Chen M, Shia BC, Wu SY. Effects of SGLT2 inhibitors on clinical cancer survival in patients with type 2 diabetes. *Diabetes Metab*. 2024;50(1):101500. doi:10.1016/j.diabet.2023.101500

9. Sung HL, Hung CY, Tung YC, Lin CC, Tsai TH, Huang KH. Comparison between sodium-glucose cotransporter 2 inhibitors and dipeptidyl peptidase 4 inhibitors on the risk of incident cancer in patients with diabetes mellitus: A real-world evidence study. *Diabetes Metab Res Rev*. 2024;40(3):e3784. doi:10.1002/dmrr.3784

10. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229-263. doi:10.3322/caac.21834

11. Hossain MS, Karuniawati H, Jairoun AA, et al. Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. *Cancers*. 2022;14(7):1732. doi:10.3390/cancers14071732

12. De Bruijn KMJ, Arends LR, Hansen BE, Leeflang S, Ruiter R, van Eijck CHJ. Systematic review and meta-analysis of the association between diabetes mellitus and incidence and mortality in breast and colorectal cancer. *Br J Surg*. 2013;100(11):1421-1429. doi:10.1002/bjs.9229

13. Pottegård A, Friis S, Christensen R dePont, Habel LA, Gagne JJ, Hallas J. Identification of Associations Between Prescribed Medications and Cancer: A Nationwide Screening Study. *eBioMedicine*. 2016;7:73-79. doi:10.1016/j.ebiom.2016.03.018

14. Pottegård A, Hallas J. New use of prescription drugs prior to a cancer diagnosis. *Pharmacoepidemiol Drug Saf*. 2017;26(2):223-227. doi:10.1002/pds.4145

15. Herweijer E, Schwamborn K, Bollaerts K, et al. Evaluation of Heterologous Effects of Travel Vaccines in Colorectal Cancer: A Database Study and a Cautionary Tale. *Gastro Hep Adv*. 2022;1(4):531-537. doi:10.1016/j.gastha.2022.02.013

16. Rytgaard HCW, van der Laan MJ. Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Anal*. 2024;30(1):4-33. doi:10.1007/s10985-022-09576-2

17. Lappe V, Köster I, Schubert I. Antidiabetische Medikation in den ersten vier Therapiejahren. Eine Studie auf Basis von Krankenkassendaten. *DMW - Dtsch Med Wochenschr*. 2017;142(01):e1-e9. doi:10.1055/s-0042-120111

12. Appendices

Code list: Appendix_codelist.xlsx