Observational Study Protocol MB102-118 ST

Comparison of the Risk of Cancer Between Patients With Type 2 Diabetes Exposed to Dapagliflozin and Those Exposed to Other Antidiabetic Treatments

AZ Study Director
Robert J. LoCasale, PhD, MS
[redacted]

External Investigators
Lia Gutierrez, BSc, MPH
RTI Health Solutions
[redacted]

Stephan Lanes, PhD, MPH; Principal Epidemiologist
Safety and Epidemiology; HealthCore, Inc.
[redacted]

PHARMO Institute, Coinvestigator
[redacted]

AstraZeneca Pharmaceuticals LP
[redacted]

SYNOPSIS

Observational Study Protocol MB102-118 ST

Protocol Title: Comparison of the Risk of Cancer Between Patients With Type 2 Diabetes Exposed to Dapagliflozin and Those Exposed to Other Antidiabetic Treatment

Department: AstraZeneca Global Epidemiology

Objectives: The ***primary objectives*** of this study are (1) to compare the incidence of breast cancer, by insulin use at cohort entry (the index date), among females with type 2 diabetes who are new users of dapagliflozin and females who are new users of antidiabetic drugs (ADs) in classes other than sodium-glucose cotransporter 2 (SGLT2) inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy and (2) to compare the incidence of bladder cancer, by insulin use at the index date and pioglitazone use, among male and female patients with type 2 diabetes who are new users of dapagliflozin and those who are new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy. ***Secondary objectives*** will compare, by insulin use at the index date, frequency of several measures of health care use, baseline characteristics, and incidence of selected other cancers in males and females between the two exposure cohorts.

Study Design: This will be a multinational cohort database study that will be conducted with data from the Clinical Practice Research Datalink (CPRD) in the United Kingdom (UK), the PHARMO Database Research Network (PHARMO) in the Netherlands, and the HealthCore Integrated Research Database (HIRD^{SM}) and Medicare database in the United States of America (US). The study will compare the incidences of certain cancers among new users of dapagliflozin with those among new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy. The planned study duration is 10 years; actual duration will depend on the actual use of dapagliflozin in the populations covered by the targeted health data sources.

Study Population: Eligible patients must meet ***all*** of the following ***inclusion criteria***: (1) patient was newly prescribed dapagliflozin or newly prescribed an AD (with or without other ADs) in a class other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy on the index date; (2) patient is aged 40 years or older at the index date; and (3) patient was enrolled in the data source for at least 180 days before the index date. Eligible patients must have ***none*** of the following ***exclusion criteria***: (1) any recording of a previous prescription of a non-dapagliflozin SGLT2 inhibitor on or before the index date; (2) any evidence of diagnosis of type 1 diabetes before the index date or use of insulin alone as the first recorded AD; (3) any diagnosis of invasive cancer before the index date (other than nonmelanoma skin cancer); (4) for the bladder cancer cohort only, any recording of hematuria, cystoscopy or urine cytology performed within 6 months before the index date; and (5) for the breast cancer cohort only, any breast biopsy performed within 6 months before the index date. All dates of new use of dapagliflozin or eligible comparator ADs occurring during the study period will be evaluated as possible index dates. In the HIRD^{SM} and PHARMO databases, all available eligible comparator patients will be used in the analysis. Depending on data availability, all eligible comparator patients will also be used in the analysis in the CPRD and in Medicare. However, if access to all comparator patients cannot be obtained, comparator patients will be frequency matched to the dapagliflozin patients in at least a 6:1 ratio (CPRD) or at least a 15:1 ratio (Medicare) by 5-year age groups, sex, geographic region, and calendar year of the index date. Dapagliflozin exposure is of primary interest to the study objectives, therefore all eligible dapagliflozin initiators will enter the dapagliflozin-exposed cohort, even if an eligible AD comparator is initiated simultaneously with or subsequently to dapagliflozin. Only the first qualifying AD comparator episode will be considered eligible for matching and entry of a patient into the comparator cohort.

Data Collection Methods

Data Sources: The study will be conducted as a multinational study in populations covered in four population-based automated health databases: the CPRD in the UK, which contains electronic medical records including outpatient diagnoses and prescriptions from general practitioner practices and mentions of diagnoses associated with hospitalizations; the PHARMO Database Network in the Netherlands, which includes health databases linked on a patient level—including community (outpatient) pharmacy data and hospitalization data, the Dutch National Pathology Registry (PALGA), and the Netherlands Cancer Registry; and, in the US, the HIRD^{SM}, which contains health insurance claims from one of the largest commercially insured population in the US, and the Centers for Medicare and Medicaid Services (CMS) Medicare data, which include health insurance claims from the federally sponsored health insurance program for individuals in the US aged 65 years or older and individuals with permanent disabilities.

Exposures: New use of dapagliflozin will be defined as the date of first dapagliflozin prescription in the data source (index date). New use of an AD in a different allowed AD class will similarly be defined as the date of first prescription for such a medication in the data source (index date).

Outcomes: The primary outcomes in this study are female invasive breast cancer and in situ and invasive bladder cancer in males and females. Secondary outcomes, not all of which are available in all data sources, include the frequency of health service utilization including the number of physician, emergency department, and hospital visits; the number of specialty care visits; and numbers of urine cytologies, cystoscopies, visits for hematuria, mammograms, and breast biopsies. In addition, secondary invasive cancer outcomes include separate composite endpoints for males (prostate, colon/rectum, lung, stomach, non-Hodgkin lymphoma [NHL], and melanoma of skin) and for females (colon/rectum, lung, corpus uteri, ovary, stomach, NHL, and melanoma of skin).

Follow-up: The date of cohort entry (the index date) will be the date an eligible patient has a first prescription or dispensing for dapagliflozin or another eligible AD class. Patients in the dapagliflozin cohort will remain in the dapagliflozin-exposed cohort, even if patients initiate treatment with other allowed ADs. However, a patient originally selected for entry into the comparator cohort might subsequently become eligible to enter the dapagliflozin-exposed cohort, if dapagliflozin is initiated. Follow-up will begin on the day after, but not including, the index date and will continue until the first occurrence of any of the cancer endpoints, death, discontinuation from the study database, or the end of the study. In a sensitivity analysis, follow-up will continue until the first occurrence of each specific cancer endpoint (regardless of a prior occurrence of any cancer during the study period).

Data Analyses: Descriptive statistics will be calculated to compare baseline characteristics (e.g., demographic information, comorbidities, and medication use) at the index date between dapagliflozin initiators versus comparator AD initiators, by insulin use at the index date. Propensity scores at the index date will be estimated by logistic regression analyses, incorporating measured potential predictors of exposure group and calendar year of the index date as independent variables in the regression model and actual exposure group (dapagliflozin or comparator) as the outcome. Duration of lookback time will be included in the model. Incidence rates of female breast cancer, bladder cancer in men and women, and the two composite cancer outcomes (one in men and the other in women) will be determined in each cohort. Propensity score–stratified analysis will be used to estimate adjusted incidence rate ratios (IRRs) of the outcomes of interest with 95% confidence intervals in dapagliflozin initiators versus other AD initiators. Changes in the intensification of diabetes treatment (based on the number of antidiabetic drugs of different classes) during follow-up will be evaluated as a time-dependent variable. In addition, an exploratory analysis will be conducted to assess if there is a trend of increasing cancer risk with increase in diabetes severity during follow-up. If there is a substantial association, diabetes severity will be evaluated via time-dependent Cox proportional hazards regression. Analyses will be conducted in each data source, and a pooled estimate will be calculated if deemed appropriate.

Sample Size/Power: The observed study size will depend upon the market uptake of dapagliflozin in the populations covered by each of the study data sources. It is expected that approximately 80% of the accrued person-years will be contributed by individuals not on insulin at the index date and 20% by those on insulin at the index date. Based on several assumptions, we estimate that over 10 years, there will be 9,500 person-years of dapagliflozin-exposed follow-up (7,600 person-years from those not on insulin at the index date) available in the CPRD and 5,800 person-years of dapagliflozin-exposed follow-up (4,640 person-years from those not on insulin at the index date) available in PHARMO databases.

In the US, based on several assumptions, we estimate that there will be approximately 835,000 person-years of follow-up available among all new users of dapagliflozin (138,000 person-years in the HIRD^{SM} and 697,000 person-years in Medicare data) over 9 years. This exposure would include approximately 668,000 person-years among those not on insulin at the index date and 167,000 person-years among those on insulin at the index date. If females contribute half of the person-time, we expect 55,000 exposed person-years for females not on insulin at the index date in the HIRD^{SM} and 279,000 person-years in Medicare data.

Limitations/Strengths: In the CPRD, there may be inaccuracies in the recorded dates of cancer diagnosis and missing information from specialists. In the PHARMO Database Network, access to clinical information will include hospital discharge diagnoses, outpatient prescription dispensing data, pathology diagnoses, clinical laboratory data, cancer registry data, and general practitioner data for a subcohort of the included patients. In the HIRD^{SM} and Medicare cohorts, the health insurance claims databases include claims for all medical services for cohort members during the study period. The Medicare data cover a very large proportion of US residents aged 65 years or older, and the HIRD^{SM} covers a large proportion of the US population younger than 65 years of age. Information on potentially important confounders such as high body mass index and smoking is virtually nonexistent unless treatment for either is detectable through claims. Therefore, an evaluation of the impact of missing confounders is planned.

Differences in the availability of data to identify confounding variables and medical records across data sources, misclassification of exposures and outcomes, and the expected relatively small number of available dapagliflozin-exposed study patients are additional limitations of this study**.** Detection bias and channeling bias are of specific concern for this study. Given dapagliflozin’s mechanism of action and potential labeling for elevated risks of breast and bladder cancer, if dapagliflozin-exposed patients are subjected to increased medical surveillance and cancer diagnostic procedures, IRRs for these cancers may be biased upward during the early period of follow-up after treatment initiation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Dapagliflozin (BMS-512148) is a highly potent, selective, and reversible inhibitor of the human renal sodium-glucose cotransporter 2 (SGLT2), the major transporter responsible for renal glucose reabsorption. Dapagliflozin lowers plasma glucose by inhibiting the renal reabsorption of glucose and by promoting its urinary excretion, making it a member of an emerging therapeutic class in the treatment of type 2 diabetes mellitus (T2DM) (Bristol-Myers Squibb [BMS] and AstraZeneca [AZ], 2011).

Diabetes has been associated with an increased risk of several cancers, including those of the pancreas, liver, breast, colon and rectum, urinary tract, and female reproductive organs (Vigneri et al., 2009), and a decreased risk of prostate cancer (Kasper and Giovannucci, 2006; Kasper et al., 2009). The assessment of possible effects of antidiabetic drugs (ADs) on the risk of cancer should be considered against this background because any such effects may be superimposed upon direct or indirect effects of diabetes itself (and/or the prediabetic phase of the disease). Changes in cancer risk associated with diabetes and its treatment have been hypothesized to be mediated through several possible pathophysiologic mechanisms including hyperglycemia, hyperinsulinemia, and the action of insulin-like growth factor 1 and may be related at least in part to the prediabetic state characterized by increasing insulin resistance and hyperinsulinemia (Giovannucci et al., 2010).

Possible effects of ADs on cancer risk may be confounded by the higher prevalence among patients with diabetes than the general population of risk factors for diabetes that are also risk factors for cancer. Such risk factors include obesity, decreased physical activity, energy-dense diet, alcohol use, and smoking. For example, in a systematic review and meta-analysis of 25 prospective observational cohort studies, 24 reported an association between active smoking and the incidence of T2DM (pooled adjusted relative risk [RR], 1.44; 95% confidence interval [CI], 1.31-1.58) (Willi et al., 2007).

Short-term incidence of cancer among patients with newly diagnosed T2DM could also be affected by surveillance bias in a population undergoing frequent medical evaluation in whom diagnostic testing may be prompted by adverse effects of ADs. For example, if urinary tract infection or other urinary symptoms occur more commonly in patients treated with dapagliflozin than in the comparator cohort, it is possible that more cystoscopies or urine cytology examinations may be performed in the dapagliflozin-treated group.

Epidemiology of Breast Cancer Among Patients With T2DM

Several studies have shown an increased risk in breast cancer among women with diabetes. A meta-analysis of published studies conducted in populations of nine countries (Canada, Denmark, Italy, Japan, Korea, the Netherlands, Sweden, the United Kingdom, and the United States) found that women with diabetes had approximately a 20% higher risk of breast cancer than women without diabetes (RR, 1.20; 95% CI, 1.12-1.28) (Larsson et al., 2007). In the Nurses' Health Study, a modestly elevated risk of breast cancer (hazard ratio [HR], 1.17; 95% CI, 1.01-1.35) was observed in patients with diabetes regardless of family history, age, obesity, physical activity, alcohol consumption, and reproductive factors (Michels et al., 2003).

Treatment with ADs may affect a diabetic woman's risk of breast cancer. In a case-control study conducted in a population of diabetic women in the General Practice Research Database, long-term use of metformin (≥ 40 prescriptions, corresponding approximately to 5 or more years) was associated with a decreased risk of breast cancer compared with nonuse (adjusted odds ratio, 0.44; 95% CI, 0.24-0.82) (Bodmer et al., 2010). Similarly, in a report from the Women's Health Initiative, metformin users among postmenopausal women with diabetes had a lower risk of breast cancer than postmenopausal women without diabetes (HR, 0.75; 95% CI, 0.57-0.99), whereas users of other antidiabetic drugs had a higher risk (HR, 1.16; 95% CI, 0.93-1.45) (Chlebowski et al., 2012). By contrast, in a study of patients with T2DM using data from The Health Information Network, no material differences were found in breast cancer risk among women treated with metformin compared with those who received sulfonylureas, those who received both metformin or sulfonylureas, or those who received insulin-based therapies. (Currie et al., 2009). Also, in a study carried out in the PHARMO Database Network, a lower risk of all malignancies was reported among patients treated with insulin glargine compared with patients treated with human insulin (HR, 0.75; 95% CI, 0.71-0.80), but the risk of breast cancer in the insulin glargine–treated group was reported to be increased (HR, 1.58; 95% CI, 1.22-2.05) (Ruiter et al., 2012).

Epidemiology of Bladder Cancer Among Patients With T2DM

An elevated risk of bladder cancer among patients with diabetes has been reported in several observational studies. In a study by Bristol-Myers Squibb (BMS) using the PharMetrics insurance claims database, the estimated incidence rate of bladder cancer among adults with T2DM was 55.1 per 100,000 patient-years, corresponding to a relative risk of 2.8 (95% CI, 2.6-2.9) compared with patients without diabetes (BMS, Epidemiology of Bladder Cancer in a Cohort of Adult Diabetics CV168-052, data on file, 2005). A meta-analysis of 16 studies (7 case-control studies and 9 cohort studies) found that the summary relative risk of bladder cancer among patients with diabetes compared with patients without diabetes was 1.24 (95% CI, 1.08-1.42) (Larsson et al., 2006).

Recent observational studies of pioglitazone use have suggested an increased risk of bladder cancer compared with the risk in nonusers with diabetes. In a longitudinal cohort study performed within the Kaiser Permanente Northern California diabetes registry, pioglitazone was reported to increase the risk of bladder cancer after more than 24 months of treatment (HR, 1.4; 95% CI, 1.03-2.0) (Lewis et al., 2011). These results were consistent with those from a retrospective cohort study conducted within 1,491,060 patients with T2DM aged 40-79 years followed for 4 years (2006 to 2009) in a large French health insurance organization (Neumann et al., 2012). Overall, 2,016 bladder malignancies were diagnosed, including 175 among 155,535 pioglitazone users. An elevated risk of bladder cancer was found for men compared with women (HR, 7.7; 95% CI, 6.7-8.8) and for patients exposed to pioglitazone compared with unexposed patients (HR, 1.22; 95% CI, 1.05-1.43). More precisely, the increased risk was found in men, after 360 days of exposure, and after a cumulative dose of pioglitazone of at least 28,000 mg.

1.1 Study Rationale

Cancer Data From Premarketing Clinical Trials

As of May 2011, breast cancer was reported in 10 female patients (9 on dapagliflozin and 1 on control) across 17 completed phase 2b and 3 studies in the dapagliflozin clinical program. Female patients treated with dapagliflozin experienced a 0.4% risk of breast cancer versus 0.1% of controls; the incidence rate ratio was 4.41 (95% CI, 0.57-200.86). The estimated rate difference compared with controls was 339 events per 100,000 patient-years (95% CI, –381 to 899), corresponding to detection of 1 excess case per 295 patient-years. All patients with breast cancer were aged more than 50 years, and 8 of the 10 patients were aged more than 60 years. Seven patients were also treated with other antidiabetic medications: insulin (n = 3), metformin (n = 3), and glimepiride (n = 1). All except a 53-year-old subject were postmenopausal. All cases were detected less than 1 year after exposure to dapagliflozin, and two were reported within the first 8 weeks of treatment. This short duration of exposure is inconsistent with the latency period for the development of chemically induced human breast cancers, which is typically several years to decades (Malone, 1993). Patients with breast cancer came from nine different countries across three continents, indicating no geographic clustering of the events.

Since the time of the first Advisory Committee meeting, there have been three more cases of breast cancer on dapagliflozin and two more cases on control (BMS and AZ, 2013). The total number of breast cancer cases on dapagliflozin is 12 (0.45%), with an exposure adjusted incidence rate of 0.40 (95% CI, 0.21-0.70) vs. 3 cases on control, with an exposure adjusted incidence rate of 0.19 (95% CI, 0.04-0.56). The incidence rate ratio is 2.47 (95% CI, 0.64-14.10). With the new cases, the characteristics of the breast cancer cases continue to reflect those seen in the general population with respect to patient age and sex and with respect to tumor heterogeneity; as before, most of the cases were diagnosed within 1 year of treatment initiation, a short time frame for carcinogenesis.

As of May 2011, bladder cancer was reported in 10 male patients (9 on dapagliflozin and 1 on control) across 19 phase 2b and 3 studies. The risk of bladder cancer was 0.06% among men treated with dapagliflozin versus 0.03% among those in the control group. The estimated incidence rate difference compared with controls was 125 events per 100,000 patient-years (95% CI, –180 to 376), corresponding to detection of 1 excess case per 800 patient-years. All patients with bladder cancer were male and most were aged 60 or more years. Microscopic or trace hematuria was reported for six patients before study treatment with dapagliflozin or placebo, which may indicate the presence of pre-existing bladder cancer (Kirkali et al., 2005). Six patients were also treated with other antidiabetic medications: insulin (n = 3), metformin (n = 2), and pioglitazone (n = 1). All 10 cases were reported within 2 years of starting study treatment, with a median time to event of 393 days and a range of 43 to 727 days. The long latency period (18 to 44 years) associated with carcinogen-induced bladder cancer (Matanoski and Elliott, 1981) suggests that the possibility of dapagliflozin treatment leading to de novo cases of bladder cancer is unlikely. Patients came from eight different countries across four continents, indicating no geographic clustering of the events.

Since the integrated database lock for Study 30-MU, one subsequent additional case detected early in the treatment course has been reported in a female patient in the ongoing add-on to sulfonylurea and metformin study (BMS and AZ, 2013).

As of December 2013, there continues to be no overall imbalance in malignancies (BMS and AZ, 2013). As expected for a drug that does not cause cancer, variability in incidence rates across the different types of cancer continues to result in a number of organ systems where the malignancy incidence rate is lower in the control group and a number of organ systems where the incidence rate is lower in the dapagliflozin group. As before, none of the imbalances are statistically significant.

Whereas numerical imbalances were observed in the incidence of breast and bladder cancer between dapagliflozin-treated patients and controls, the overall proportions of malignant and unspecified tumors in phase 2b and 3 studies were 1.4% of dapagliflozin-treated patients versus 1.3% of control patients. In addition to breast and bladder cancer, we will also estimate the incidence of several additional cancer types in this study. To select cancers for study, we focused our attention on the 10 leading cancers by incidence in the European Union in 2012, as reported by Ferlay et al. (2013): prostate, female breast, large bowel, lung (including trachea and bronchus), corpus uteri, bladder, malignant melanoma of skin, ovary, kidney (including renal pelvis and ureter), and non-Hodgkin lymphoma (NHL). Two of these cancers (breast and bladder) are already specified as distinct endpoints in this study. Three others (prostate, corpus uteri, and ovary) are sex-specific, making it problematic to combine them into a composite cancer endpoint since not all cohort members would be at risk for all components of the endpoint. Therefore, we will evaluate two additional secondary endpoints, one in males and the other in females, defined as the composite incidence rate for the leading cancers (other than breast and bladder) for which males and females, separately, are at risk.

This postauthorization safety study (PASS) is being conducted as part of the BMS/AstraZeneca (AZ) Dapagliflozin Risk Management Plan to monitor the safety of dapagliflozin in real-world use. This study is complementary to a proposed large cardiovascular outcome clinical trial where the risk of breast and bladder cancer will also be evaluated. This PASS will provide insight regarding the demographics of patients using dapagliflozin in usual clinical practice and is designed to estimate cancer risk among patients using dapagliflozin.

As in most observational studies, the results of this PASS may be affected by detection bias or by channeling bias. Given dapagliflozin’s mechanism of action and potential labeling for breast and bladder cancer, if dapagliflozin-exposed patients are subjected to increased medical surveillance and cancer diagnostic procedures, the hazard ratios for breast and bladder cancer in the dapagliflozin cohort compared with users of other ADs may be biased upward. Channeling bias could affect the study in two ways: (1) dapagliflozin could be preferentially prescribed to patients with fewer risk factors for breast and bladder cancer, thereby biasing the hazard ratio downward, or (2) dapagliflozin could be preferentially prescribed to patients with more severe diabetes (or to patients who have failed other therapies), potentially with more risk factors for the outcomes, thereby biasing the hazard ratio upward. Efforts to document and (where possible) address these methodological issues are discussed in this protocol.

1.2 Research Questions

Research question 1: What is the estimated risk of breast, bladder, and other common cancers for patients with T2DM who are new users of dapagliflozin compared with those who are new users of antidiabetic treatments (ADs) in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy?

Research question 2: Given dapagliflozin’s mechanism of action and potential labeling for breast and/or bladder cancer, is there differential medical surveillance (detection bias) for the diagnosis of these cancers for patients with T2DM who are new users of dapagliflozin compared with those who are new users of antidiabetic treatments in the other AD classes under study?

2 STUDY OBJECTIVES

2.1 Primary Objectives

Primary objective #1: To compare the incidence of breast cancer, by insulin use at the index date, among females with T2DM who are new users of dapagliflozin and females who are new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy. A list of the ADs eligible for inclusion in the comparator cohort can be found in Appendix 6.

Primary objective #2: To compare the overall and sex-specific incidence of bladder cancer, by insulin use at the index date and pioglitazone use, among patients with T2DM who are new users of dapagliflozin and those who are new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy.

2.2 Secondary Objectives

Secondary objective #1: To compare during follow-up the frequency of several measures of health care utilization (including outpatient visit frequencies and use of breast and bladder cancer screening and diagnostic tests), by insulin use at the index date, among patients with T2DM who are new users of dapagliflozin and those who are new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy.

Secondary objective #2: To compare baseline patient characteristics, by insulin use at the index date, among patients with T2DM who are new users of dapagliflozin and those who are new users of other ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy and to identify important prognostic variables that should be balanced between exposure groups and that should be included in the propensity scores used in the primary analyses.

Secondary objective #3: To compare the composite incidence of selected cancers (prostate, colon/rectum, lung, stomach, NHL, and melanoma of skin), by insulin use at the index date, among males with T2DM who are users of dapagliflozin and those who are users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy.

Secondary objective #4: To compare the composite incidence of selected cancers (colon/rectum, lung, corpus uteri, ovary, stomach, NHL, and melanoma of skin), by insulin use at the index date, among females with T2DM who are users of dapagliflozin and those who are users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy.

2.3 Exploratory Objective

Not applicable.

3 STUDY DESIGN

3.1 Overview of Study Design

This is a multinational cohort study that will use existing automated population-based and administrative health care databases in the United Kingdom (UK), the Netherlands, and the United States (US). A cohort design will allow direct estimation of the incidence and relative risk of the outcomes of interest that are potentially associated with dapagliflozin use compared with use of other ADs. Further, the cohort design permits determination of outcomes at multiple time points, as well as the assessment of risk for a variety of exposure measures.

The planned study duration is 10 years; however, duration will depend on the actual use of dapagliflozin in the populations covered by the targeted data sources.

3.2 Study Population

During the conduct of the study, patients will be identified at selected intervals, planned to be every 24 months. Study populations of patients with T2DM will be identified using data on general practice diagnoses and prescriptions in the Clinical Practice Research Datalink (CPRD) in the UK, pharmacy dispensings in the PHARMO Database Network in the Netherlands, and health insurance claims for outpatient medication dispensings in the HealthCore Integrated Research Database (HIRD^{SM}) and Centers for Medicare and Medicaid Services (CMS) Medicare databases of the US. These patients will be new users of dapagliflozin or other selected ADs, as detailed in Section 3.2.2, Inclusion Criteria.

3.2.1 Definition of the Index Date/Cohort Entry Date

Patients will enter the study cohort on the index date (see Section 3.2.2, Inclusion Criteria), which is defined as the date a patient was newly prescribed or dispensed dapagliflozin or an AD in another allowed class after the beginning of the study observation period, which is defined according to the time of approval of dapagliflozin in each country (see Section 2, Study Objectives). In the CPRD, the index date corresponds to the date a prescription is written by a general practitioner; in the PHARMO Database Network, HIRD^{SM}, and Medicare databases, the index date corresponds to the date a prescription is dispensed at a community pharmacy. All dates of new use of dapagliflozin or eligible comparator ADs occurring during the study period will be evaluated as possible index dates. Dapagliflozin exposure is of primary interest to the study objectives, therefore all eligible dapagliflozin new use dates will qualify for entry into the dapagliflozin-exposed cohort, even if an eligible AD comparator is initiated simultaneously or subsequently. Only the first qualifying AD comparator prescription episode will be considered eligible for matching and entry of a patient into the comparator cohort. A subsequent new AD prescription of a different qualifying comparator drug will not be eligible for inclusion into the comparator population for propensity score estimation. Furthermore, a patient who is selected as a comparison patient and later initiates dapagliflozin will be entered as a dapagliflozin new user and will have the comparator AD follow-up time censored and dapagliflozin follow-up time started at the time of dapagliflozin treatment initiation.

Cohort entry (the index date) is defined as the date of the first prescription or dispensing of dapagliflozin or an eligible comparator AD. On the index date, eligibility will be assessed, and patient characteristics will be recorded; propensity scores will be estimated including calendar year of the index date.

3.2.2 Inclusion Criteria

Eligible patients must meet ***all*** of the following ***inclusion criteria***:

- Patient was newly prescribed or dispensed dapagliflozin (with or without other ADs) or newly prescribed or dispensed an AD (with or without other ADs) in a class other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy on the index date.
- Patient was aged 40 years or older at the index date.
  - In the HIRD^{SM}, was aged 40-64 years
  - In Medicare, was aged 65 years or older; was a participant only in the fee-for-service program (i.e., was not in a managed care program); was enrolled in Parts A, B, and D of the Medicare program for at least 180 days before entering the study (follow-up will be censored if Part D coverage is discontinued); was a resident in a US state or the District of Columbia; and eligibility for Medicare was not due to end-stage renal disease
- Patient was enrolled in the data source for at least 180 days before the prescription or dispensing index date.

3.2.3 Exclusion Criteria

Eligible patients must have ***none*** of the following ***exclusion criteria***:

- The patient was prescribed a non-dapagliflozin SGLT2 inhibitor on or before the index date.
- The patient initiated metformin or sulfonylurea as AD monotherapy at the index date.
- The patient initiated insulin monotherapy at the index date.
- The patient had evidence of type 1 diabetes on or before the index date, or the first recorded AD is insulin monotherapy.
- The patient had any diagnosis of invasive cancer on or before the index date (other than nonmelanoma skin cancer, defined as basal and squamous cell skin cancers).
- For the bladder cancer outcome only:
  - The patient had a recording of hematuria within the 6 months prior to and including the index date.

  - The patient had a cystoscopy or urine cytology performed within the 6 months prior to and including the index date.
- For the breast cancer outcome only:
  - The patient had a breast biopsy performed within the 6 months prior to and including the index date.

3.2.4 Rationale for Inclusion and Exclusion Criteria

A minimum of 180 days of recorded information must be available before the index date to allow us to identify which prescriptions represent new use of dapagliflozin or a comparison AD. The minimum required period of 180 days before the index date is based on the assumption that prescriptions for study medications that do not represent new use will be recorded at least once during this period. This minimum required period of recorded data will also increase our ability to identify patients with a history of cancer. However, it remains possible that a small number of patients with a history of cancer may not be detected regardless of the duration of information available in the data source before the index date.

The rationale for comparing new users of dapagliflozin with new users of ADs *in a class other than SGLT2 inhibitors* is to ensure that we do not miss potentially important associations that are due to the SGLT2 class after other compounds in the same class become available.

The rationale for not including new users of *insulin monotherapy* in the comparison group is to enhance the comparability of the baseline characteristics of the comparison groups. Typically, patients with T2DM who initiate insulin are older, have more severe disease (more T2DM comorbidities), and have more poorly controlled diabetes after treatment with an oral AD. If dapagliflozin is primarily used among relatively newly diagnosed diabetics, inclusion of new insulin users in the comparison cohort could make it more difficult to obtain comparable populations. However, patients already taking insulin who “add-on” either dapagliflozin or another qualifying AD will be eligible for inclusion in this study, and all analyses will be stratified by baseline insulin use.

The rationale for not including new users of *metformin monotherapy or sulfonylurea monotherapy* in the comparison group is that patients with diabetes often receive these drugs alone as initial AD treatment early in the course of the disease (e.g., UK National Institute for Health and Clinical Excellence [2015] and American Diabetes Association [2014] guidelines). It is not expected that dapagliflozin will be commonly used alone as initial AD treatment following diagnosis; therefore, comparability of the dapagliflozin and comparison cohort should be enhanced by excluding patients who are treated initially with metformin or sulfonylurea as monotherapy. In clinical practice, patients may be newly prescribed dapagliflozin or another AD with or without other ADs already prescribed as part of their regimen (i.e., patients may have new antidiabetic medications added on or they may switch agents). Therefore, we plan to include patients, regardless of whether or not they are taking other ADs (including insulin) at the time they are newly prescribed either dapagliflozin or an acceptable study AD. We will collect information on whether patients received prior antidiabetic therapy and, if so, whether the study drug that made the patient eligible for this study was “added on” or “switched to” on the index date.

3.2.5 Selection of Subjects

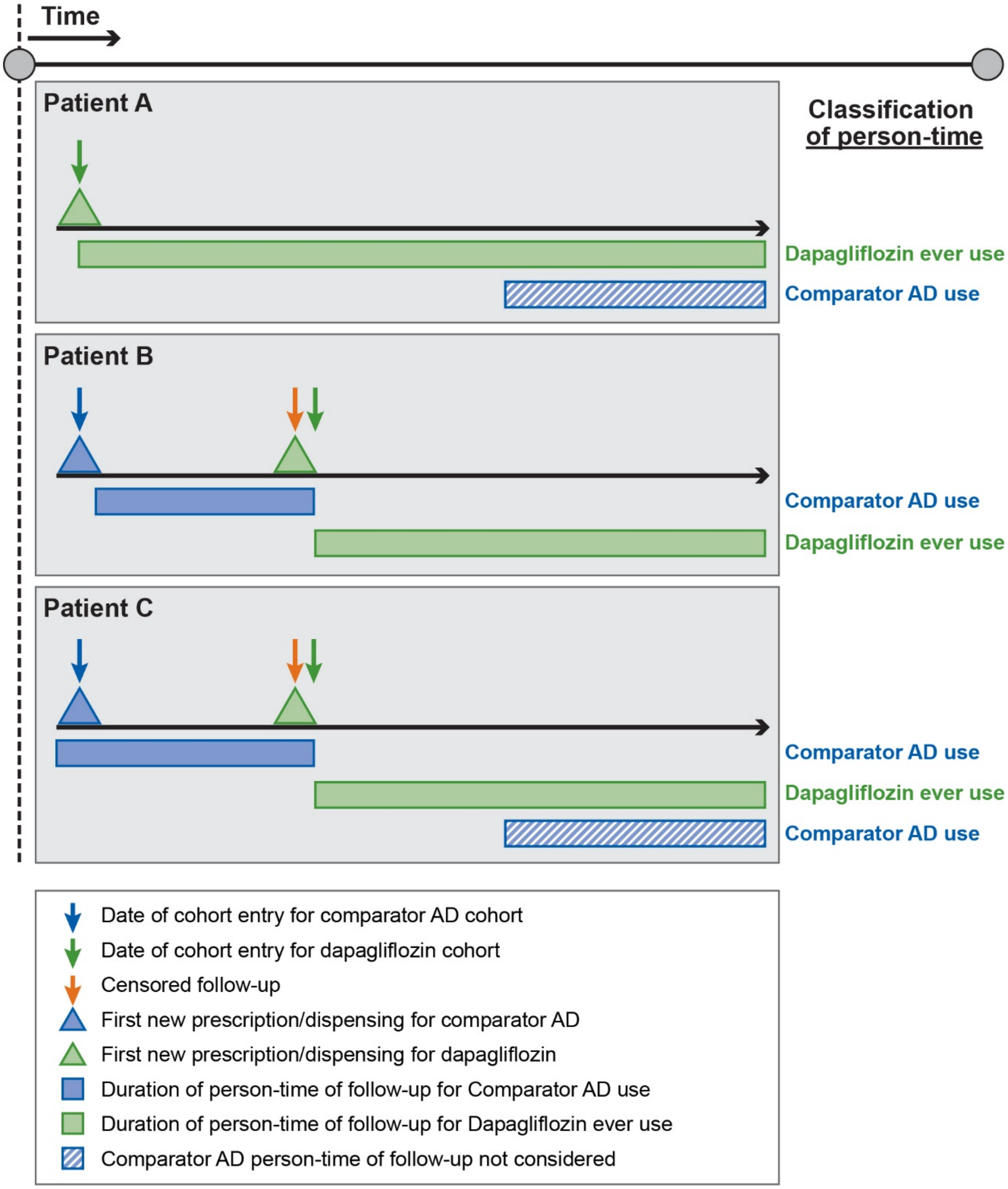
Eligible study patients will be selected from the study data sources separately. All patients in each data source who meet inclusion and exclusion criteria and are new users of dapagliflozin will be included. All patients meeting inclusion and exclusion criteria who are new users of the first allowed comparator AD will be enumerated and considered for inclusion at the time of any identified prescription that qualifies as the index prescription. In the CPRD and Medicare, it may be necessary to select a subsample of AD users before propensity score estimation. In the CPRD, the subsample will be identified by frequency matching eligible comparator patients to dapagliflozin patients in at least a 6:1 ratio by 5-year age groups, sex, geographic region, and calendar year of the index date. In Medicare, a subsample of AD users will be randomly sampled in at least a 15:1 ratio to dapagliflozin patients. In the HIRD^{SM} and PHARMO databases, all available eligible comparator patients will be used in the analysis.

If it is necessary to select a subsample in the CPRD and Medicare, it is anticipated there will be a large number of comparator AD new users in these data sources, far more than would be needed to contribute meaningful statistical power and precision. The rationale for selecting at least 6 comparator AD new users for each dapagliflozin new user in the CPRD and at least 15 comparator AD new users in Medicare is to ensure that we have sufficient numbers of patients to develop the propensity score and to conduct secondary analyses, as needed. Sampling a larger number of comparator patients for each dapagliflozin-exposed user would add limited additional statistical power and precision. Random sampling of new comparator AD users in the CPRD and Medicare will occur each time new dapagliflozin-exposed patients are drawn from each data source or when a patient previously being followed in the comparator group becomes dapagliflozin exposed and is then entered into the dapagliflozin cohort.

3.2.6 Cohort Entry and Follow-up

As shown in Figure 1, eligible patients will be entered into the study cohort on the day of their first prescription or dispensing of dapagliflozin or the first allowed comparator AD (the index date). Follow-up for any patient in either the dapagliflozin-exposed or comparator cohorts will begin on the day after, but not including, the index date and will continue until the earliest of the following events: (1) diagnosis of any cancer, (2) death, (3) discontinuation in the health care data source, or (4) end of study. If a patient develops type 1 diabetes during follow-up (physician diagnosis in the CPRD or PHARMO or fulfillment of a claims definition in the HIRD^{SM} or Medicare), follow-up time will be censored on the date of the diagnosis. In a sensitivity analysis, follow-up will continue until the first occurrence of each specific cancer endpoint (regardless of a prior occurrence of any cancer during the study period). In addition, for patients in the comparator cohort who subsequently receive dapagliflozin, follow-up time in the comparator cohort will be censored at the time they are eligible to enter the dapagliflozin cohort (i.e., the day of the prescription/dispensing of dapagliflozin will count as the last day of follow-up for the comparator AD). The rationale for this is to ensure that all new users of dapagliflozin who otherwise meet eligibility requirements are included in the dapagliflozin-exposed cohort since dapagliflozin exposure is anticipated to be the limiting factor in the study size and exposure to dapagliflozin is of primary interest for the study objectives.

Figure 1: Cohort Entry and Follow-up


Time
Patient A
Classification of person-time
Dapagliflozin ever use
Comparator AD use
Patient B
Comparator AD use
Dapagliflozin ever use
Patient C
Comparator AD use
Dapagliflozin ever use
Comparator AD use
Date of cohort entry for comparator AD cohort
Date of cohort entry for dapagliflozin cohort
Censored follow-up
First new prescription/dispensing for comparator AD
First new prescription/dispensing for dapagliflozin
Duration of person-time of follow-up for Comparator AD use
Duration of person-time of follow-up for Dapagliflozin ever use
Comparator AD person-time of follow-up not considered

AD = antidiabetic drug.

Follow-up of patients in either the dapagliflozin-exposed or comparator cohort will not end if other ADs or insulin are prescribed in addition to dapagliflozin or the comparator AD after the index date (i.e., add-on therapy). Follow-up of patients in either the dapagliflozin-exposed or comparator cohort will also not end if they switch to another AD (i.e., other than dapagliflozin). Moreover, patients in the dapagliflozin cohort who initiate treatment with other allowed comparator ADs will nonetheless continue to be followed in the dapagliflozin-exposed cohort and will not be eligible for selection into the comparator cohort because they are already considered to be dapagliflozin exposed.

At no time during the study will any patient contribute follow-up time to both the dapagliflozin-exposed and comparator cohorts simultaneously, and any cancer event will be counted only once (in the exposure category in which the patient is accumulating person-time at the time the event occurs). However, as previously noted, a patient originally selected for entry into the comparator cohort might subsequently become eligible to enter the dapagliflozin-exposed cohort. Such a patient would be counted twice in a tabulation of patients' characteristics at the index date since the patient would in effect have entered the study twice (once into each cohort). The number of such patients is expected to be low and will be reported if this occurs.

Patients whose follow-up time in the comparator cohort is censored at such time as they become eligible for entry into the dapagliflozin cohort will not retroactively be excluded from the comparator cohort because doing so would make such a patient's original eligibility for selection into the comparator cohort dependent on a future event (initiation of dapagliflozin at a later date). The person-time accumulated by such a patient up to the time he or she qualifies for entry into the dapagliflozin-exposed cohort will be included in the analysis of cancer incidence rates in the comparator cohort (i.e., the day of the prescription/dispensing of dapagliflozin will count as the last day of follow-up for the comparator AD). Therefore, such a patient will not be replaced in the comparator cohort. We expect that there will be minimal attrition of patients in the comparator cohort for this reason and that the censoring of this small amount of patient follow-up time will not have any material effect on estimation of the incidence of the study outcomes. Also, we expect that there will be some attrition of patients in both cohorts during the course of the study for the reasons listed in the first paragraph of this section; patients who leave the study for those reasons will similarly not be replaced.

3.3 Data Source/Data Collection Process

This study requires large data sources that capture longitudinal information on prescriptions (dispensing), inpatient and outpatient discharge diagnoses, and procedures on individuals and that allow for case adjudication via medical record review or linkage to a cancer registry. This study will be conducted using four longitudinal health care data sources (one in the UK, one in the Netherlands, and two in the US) that include demographic data, prescription or dispensing information, and medical diagnostic and procedure codes and/or cancer registry data.

3.3.1 Clinical Practice Research Datalink – UK

The Clinical Practice Research Datalink (previously the General Practice Research Database [GPRD]) contains diagnostic and prescribing information recorded by general practitioners (GPs) as part of their routine clinical practice in the UK. The data source coverage is approximately 4 million of the UK population. These data are linkable, at least partially, with other health care data sets (e.g., hospitalization records and national mortality data) via the patient’s National Health Service number, sex, date of birth, and postal code. Detailed information on prescriptions written by the GPs, including prescribed dosage, is automatically recorded in the data source. Read codes are used for diagnoses. Additional diagnostic and treatment information can be found in letters from specialists and hospitals, and other sources. Cancer diagnoses recorded in this data source have been found to be highly reliable in multiple studies. For example, one investigator reported that essentially all cases in the GPRD with a diagnostic code for esophageal cancer were confirmed to have had the disease (Walker, 2011). Moreover, where data were available to judge the time of clinical onset, the date was within 60 days of the date recorded in the electronic medical record in 89% of cases. Similarly, in a study of calcium channel blockers and risk of cancer, among cancer cases for whom additional information was obtained directly from the patient’s general practitioner, the diagnosis was confirmed in 95% of cases (Jick et al., 1997). In another GPRD study, changes similar to those reported in national cancer statistics were observed in age-specific breast cancer incidence patterns after the introduction of a UK national screening program (Kaye et al., 2000); although ecological, this finding provides indirect support for the validity of breast cancer diagnoses in this data source. The risk of bladder cancer has also been studied in the GPRD in relation to several exposures including acetaminophen (Kaye et al., 2001) and pioglitazone (Wei et al., 2013).

Some cases identified in the CPRD will also have information in the linked Hospital Episode Statistics (HES) database, which enables access to hospitalization data including disease and procedural coding and to cancer registry data in England. These linkages could provide validation of cancer diagnoses in patients who are subsequently hospitalized or who reside in England and are treated in a subgroup of general practices. HES-linked data will also be used to identify exclusions. Approximately 65% of the English practices contributing to the CPRD have consented to have their patient information linked, via a trusted third party, to other health care data sets via the patient’s National Health Service number, sex, date of birth, and postal code. English practices represent approximately 75% of all practices contributing to the CPRD; therefore, approximately half of the total CPRD practices have these data links.

The validation of cancer cases in the CPRD, if not possible in individual cases by review of the automated medical record entries, can be accomplished by sending questionnaires to the corresponding GPs for access to medical information related to the event of interest, including referral and hospital discharge letters, or by linkage to HES or cancer registry data. In a similar manner, information on confounding variables can be accessed from the diagnostic and other Read codes in the CPRD or questionnaires administered to the corresponding GPs.

3.3.2 PHARMO

3.3.2.1 PHARMO Database Network – the Netherlands

The PHARMO Database Network is a population-based network of health care databases and combines data from different health care settings in the Netherlands. These different data sources are linked on a patient level through validated algorithms. Detailed information on the methodology and the validation of the used record linkage method can be found elsewhere (Herings and Pedersen, 2012; van Herk-Sukel et al., 2010).

The longitudinal nature of the PHARMO Database Network system enables follow-up of more than 4 million (25%) residents of a well-defined population in the Netherlands for an average of 10 years. Data collection period, catchment area, and overlap between data sources differs. Therefore, the final cohort size for any study will depend on the data sources included. As data sources are linked on an annual basis, the average lag time of the data is 1 year. All electronic patient records in the PHARMO Database Network include information on age, sex, socioeconomic status, and mortality.

The PHARMO databases planned for use in this study are described below.

Outpatient Pharmacy Database

The Outpatient Pharmacy Database comprises health care products prescribed by GPs or specialists and dispensed by an outpatient pharmacy. The dispensing records include information on type of product, date, strength, dosage regimen, quantity, route of administration, prescriber, and costs. Drug dispensings are coded according to the World Health Organization (WHO) Anatomical Therapeutic Chemical (ATC) Classification System (www.whocc.no/atc_ddd_index). Outpatient pharmacy data cover a catchment area representing 3.6 million residents.

Hospitalization Database

Taken from the Dutch hospital data, the Hospitalization Database comprises hospital admissions for more than 24 hours and admissions for less than 24 hours for which a bed is required. The records include information on discharge diagnoses, procedures, and hospital admission and discharge dates. Diagnoses are coded according to the International Classification of Diseases, and procedures are coded according to the Dutch Classification of Procedures. For more information, see: www.dutchhospitaldata.nl.

Clinical Laboratory Database

The Clinical Laboratory Database comprises results of tests performed on clinical specimens. These laboratory tests are requested by GPs and medical specialists in order to get information concerning diagnosis, treatment, and prevention of disease. The electronic records include information on date and time of testing, test result, unit of measurement and type of clinical specimen. Laboratory tests are coded according to the Dutch WCIA coding system (https://aut.nhg.org/labcodeviewer/). Clinical laboratory data cover a catchment area representing 1.2 million residents.

General Practitioner Database

The General Practitioner (GP) Database comprises data from electronic patient records recorded by GPs. The records include information on diagnoses and symptoms, laboratory test results, referrals to specialists, and health care product/drug prescriptions. The prescription records include information on type of product, date, strength, dosage regimen, quantity, and route of administration. Drug prescriptions are coded according to the WHO ATC classification system (www.whocc.no/atc_ddd_index). Diagnoses and symptoms are coded according to the International Classification of Primary Care (https://www.nhg.org/themas/artikelen/icpc), which can be mapped to ICD (International Classification of Diseases) codes, but can also be entered as free text. GP data cover a catchment area representing 1.5 million residents.

National Pathology Registry

The nationwide network and registry of histo- and cytopathology in the Netherlands is maintained by the PALGA (Dutch National Pathology Registry) and comprises excerpts of histological, cytological, and autopsy examinations. Electronic records include information from abstracts of pathology reports, consisting of a summary of the report and the PALGA diagnosis, which is structured along five classification axes: topography, morphology, function, procedure, and diseases. To obtain these data, permission is needed on a project basis. For more information, see www.palga.nl.

3.3.2.2 Netherlands Cancer Registry

The Netherlands Cancer Registry (NCR) is maintained by the Netherlands Comprehensive Cancer Organisation and comprises information on newly diagnosed cancer patients in the Netherlands, including cancer diagnosis, tumor staging (according to the TNM-classification developed and maintained by the Union for International Cancer Control (UICC, [www.uicc.org])), tumor site (topography) and morphology (histology) (according to the WHO International Classification of Diseases for Oncology (ICD-O-3, [www.who.int])), comorbidity at diagnosis, and treatment received directly after diagnosis. To obtain NCR data, permission is needed on a project basis. For more information, see www.iknl.nl.

3.3.3 HealthCore Integrated Research Database – US

HealthCore, Inc., (hereafter, HealthCore), established in 1996, is a wholly owned subsidiary of Anthem, Inc., which is one of the largest health benefits companies in the US in terms of medical membership. Anthem is an independent licensee of the Blue Cross and Blue Shield Association and serves its members as the Blue Cross licensee in 14 states and through UniCare. Anthem is also the parent of Health Management Corporation, a preventive health and disease management company.

The HIRD^{SM} contains fully adjudicated paid claims from the largest commercially insured population in the US, with dates of service for all noncapitated ambulatory, emergency department, inpatient, and outpatient encounters (including administrative claims for laboratory tests) for members with eligibility at the time of service. It also includes claims for outpatient dispensings of prescription pharmaceuticals from pharmacies. The full HIRD^{SM} database dates back to 01 January 2006, with a subset of all the plans in the database and to 01 January 2004, for all the plans represented in the database. The majority of data can be accessed from that time period through the most recent update. Data are updated monthly, with an approximate 3-month time lag for greater than 95% capture of paid medical claims. The lag for pharmacy data is shorter, with approximately 98% paid within 30 days. As of January 2014, the HIRD^{SM} contained claims information for approximately 35.8 million lives available for research. In addition, HealthCore has the ability to abstract inpatient and outpatient medical records for the health plan members represented in the HIRD^{SM}, identify and contact providers and members for survey research through vendor relationships, and link data to national vital records. The HIRD^{SM} enables rapid access to US population-based health data resources representing all major geographic regions and health care settings and varied clinical indications that permit long-term longitudinal patient follow-up. The specific geographic regions represented in the HIRD^{SM} include the Northeast, Mid-Atlantic, Southeast, Midwest, Central US, and West. The HIRD^{SM} has been used as a data source in multiple studies related to safety outcomes and validation.

Health plans contributing data to the HIRD^{SM} include several different lines of business such as health maintenance organizations, point-of-service plans, preferred provider organizations, and indemnity plans.

Patient enrollment data, medical care, prescription drug use, laboratory test results, and health care utilization can be tracked for each patient in the database. Diagnoses and procedures are identified by ICD-9-CM or ICD-10-CM,[1] Current Procedural Terminology (CPT), and Healthcare Common Procedure Coding System (HCPCS) codes for both outpatient visits and inpatient stays. Drug claims are captured by National Drug Codes (NDCs), which can be translated to broader, more meaningful classification systems such as Generic Product Identifier codes. Standard Logical Observation Identifier Names and Codes are used to define specific laboratory test result data. Physician, specialist, and emergency department visits, as well as hospital stays, are captured in the database through CPT codes, uniform billing (UB-92) revenue codes (e.g., room and board), and place-of-service codes. Information on physician specialty is also retained in the database.

Patients 65 years of age or older will be excluded from this data source to avoid any duplication with the Medicare data source. In addition, patients in the HIRD^{SM} will be censored during follow-up the day before their 65th birthday.

3.3.4 Medicare – US

Medicare is a US federal benefit program that provides health insurance for citizens and permanent residents aged 65 years or older and some disabled people younger than 65 years. Medicare coverage comprises four parts: Part A: Hospital Insurance; Part B: Medical Insurance; Part C: Medicare Advantage; and starting in 01 January 2006, Part D: Medicare Prescription Drug Coverage (CMS, 2013).

Data for services provided under Part A, B, and D insurance are claims for payment that are submitted to Medicare by an individual provider or a health care facility. Claims are intended to record the service that was provided, using detailed diagnostic, procedure, and drug codes, for Medicare reimbursement of each claim service. The diagnosis recorded on the claim is used by Medicare to understand the justification for the service, as coverage excludes some care that is deemed not medically necessary. However, diagnoses on a claim cannot be presumed to be clinically confirmed. Distinguishing claims diagnoses that are being ruled out from those that are

[1] ICD-9-CM = *International Classification of Diseases, 9th Revision, Clinical Modification*;
ICD-10-CM = *International Classification of Diseases, 10th Revision, Clinical Modification*.

confirmed is a challenge in some claims database studies. Data on laboratory test results are not available.

Similar to the HIRD^{SM} database, diagnoses and procedures are identified by ICD-9-CM or ICD-10-CM, CPT, and HCPCS codes for both outpatient visits and inpatient stays. Additionally, the Part D data claims file contains information on prescription drug fills, including product codes (NDCs), quantity dispensed, and days' supply.

There is currently a two-year lag in accessing Medicare Part D data. Generally, Medicare releases Part D data each January. Therefore, if the first interim comparative analysis occurs in January 2018, Medicare data would be available through the end of 2015.

3.4 Definitions of Study Variables

3.4.1 Outcomes/Endpoint Variables

Primary outcomes include invasive breast cancer among females only (Appendix 1, Table 1-1; Appendix 2, Table 2-1; and Appendix 3, Table 3-1) and invasive and in situ bladder cancer among males and females (see codes in Appendix 1, Table 1-2; Appendix 2, Table 2-1; and Appendix 3, Table 3-2). Secondary invasive cancer outcomes include prostate, colon/rectum, lung, stomach, NHL, and melanoma of skin among males only and colon/rectum, lung, corpus uteri, ovary, stomach, NHL, and melanoma of skin among females only (see codes in Appendix 1, Table 1-3 through Table 1-10; Appendix 2, Table 2-1 through Table 2-3; and Appendix 3, Table 3-3 through Table 3-10).

Secondary outcomes include the individual frequencies of measures of health service utilization such as the number of physician, emergency department, hospital, and specialty care visits; urine cytology urinalysis (including hematuria); cystoscopy; mammography; and breast biopsy (see codes in Appendix 4).

3.4.1.1 Electronic Case Identification

The approach for case identification will be tailored for each data source based on the characteristics of the data source and prior knowledge related to the ascertainment and validation of cancer endpoints. All cases initially identified will be considered provisional cases and at least a sample will undergo case validation.

- In the CPRD, clinical Read codes, which are the standard clinical terminology system used in general practice in the UK, and ICD-10[2] codes (for patients in the general practices able to link to HES data) will be used to screen for potential cancer cases in the study cohorts. Recording of at least one Read code or one ICD-10 code for a targeted neoplasm in a patient included in the study cohorts will be considered sufficient to meet provisional case designation (see Appendix 1 for Read codes).
- In the PHARMO Database Network, *International Classification of Diseases for Oncology, Third Edition* (ICD-O-3) codes will be used in the linked cancer registry data to identify all cancer cases of interest for the study. Secondary outcomes data will be available in the PALGA database: urine cytologies (PALGA code: T7X100 with type of

[2] ICD-10 = *International Statistical Classification of Diseases and Related Health Problems, 10th Revision.*

research performed = cytology), bladder biopsies (PALGA code: T74XXX and P114X with type of research performed = histology), and breast biopsies (PALGA code: T04XXX and P114X with type of research performed = histology). The PALGA regisstry contains information on the entire population of the Netherlands and can be linked to all patients in the other data sets being used in this study.

- In the HIRD^{SM} and Medicare databases, initial case identification will be based on the application of algorithms that have been applied in previous research. Details on the algorithms will be described in the statistical analysis plan. For assessment of secondary outcomes, ICD-9-CM procedural codes will be used and will be specified in the statistical analysis plan.

Final code lists will be included in the data development plan to be developed once the protocol is final. Other health care utilization variables, such as ADs prescribed after the index date, will also be explored. Some of these outcomes may be assessable only in the CPRD, HIRD^{SM}, and Medicare data because primary care records are not available for most of the cohort to be included from PHARMO. Final code lists for diagnostic and procedure codes for secondary outcomes will be included in the statistical analysis plan.

If possible, stage of the tumor at the time of diagnosis will be ascertained for breast and bladder cancer cases from the data sources and/or from the medical record.

3.4.1.2 Case Validation for Breast and Bladder Cancer Cases

For each data source, the validation process will be detailed in the validation plan, to be developed in the future. In summary, patients initially identified by clinical Read codes (Appendix 1) in the CPRD and by ICD-9-CM and ICD-10-CM codes (Appendix 3) in the HIRD^{SM} and Medicare data with primary malignancies of interest will be considered provisional cases. For a subset of up to 125 provisional cases with bladder cancer and up to 125 provisional cases of female breast cancer, we will conduct further validation, blinded to study drug exposure, via electronic medical record review (CPRD clinical file) or linkage of data from GP questionnaires to HES or cancer registry data in the CPRD, and by medical record review in the HIRD^{SM} and Medicare data. Additional case validation in PHARMO will not be necessary as all cases will be identified through linked cancer registry data.

Additional detail for each data source is provided below.

CPRD

In the CPRD, provisional cases will be considered confirmed if there is supportive evidence of a cancer diagnosis, specifically, a relevant pathology (morphology) Read or ICD-10 code or evidence of appropriate cancer-specific therapy (surgery, radiation therapy, chemotherapy, hormonal therapy, or other targeted or biological therapy), or a code for cancer care review, within the period from 1 month before to 3 months after at least two recorded diagnostic codes for the endpoint malignancy or, in the case of the cancer care review code, at any time after the diagnoses (but with no additional, different cancer diagnosis in the interval between the study cancer diagnosis and the occurrence of the cancer care review code). Provisional cases will also be considered confirmed if subsequent clinical events (referrals, hospitalizations, or death) are associated with appropriate clinical Read or ICD-10 codes for the cancer diagnosis. Mastectomy will be considered sufficient evidence of a diagnosis of breast cancer, but excisional biopsy (lumpectomy) will not be considered sufficient because it can be used as either a diagnostic or therapeutic procedure and therefore some excisional biopsy specimens show no evidence of malignancy. Similarly, cystectomy within 1 month before or 3 months after a recorded code of bladder cancer will be considered sufficient evidence of a diagnosis of bladder cancer, but transurethral resection alone will not be considered sufficient evidence because it can be used as either a diagnostic or a therapeutic procedure.

If after review of additional information in the electronic medical records (CPRD clinical file) it is not possible to decide whether a provisional case is a confirmed case or is not a case, consideration will be given to attempting further validation using (1) questionnaires to GPs (which can be conveyed by staff at the CPRD who will mask any personally identifying information before forwarding the information received to the analysts) or (2) linkage with the cancer registry in England or HES data. Currently, approximately 65% of the English practices contributing to the CPRD have consented to have their patient information linked, via a trusted third party, to other health care datasets via the patient’s National Health Service number, sex, date of birth, and postal code. English practices altogether represent approximately 75% of practices contributing to the CPRD; therefore, approximately half of the total CPRD practices have this link. (It is expected that as the CPRD expands over the next 2 years there will be a further increase in the proportion of the covered population who have linked information in these external data sources.) Such additional case validation will take additional time to complete beyond the time needed for electronic medical record review. Therefore, at the time of each analysis, there will likely be some confirmed cases and some cases that remain classified as provisional cases. At subsequent analyses, depending on what additional data are available, a provisional case may be found to be a confirmed case, may be found not to be a case, or may remain a provisional case if insufficient additional information has become available since the previous analysis. Analyses to address the study objectives will be conducted using cases identified using the electronic coding algorithm that was validated during the medical record review; separate analyses will be conducted using only confirmed cases. The primary analysis will have some misclassification, but if the positive predictive value of the algorithm is relatively high, the results of the two analyses will be similar, and estimates from the combined analysis will be more precise than estimates from the analysis restricted to confirmed cases.

PHARMO

Additional procedures for case validation in PHARMO data will not be necessary in this study because cases will be ascertained through linkage to the NCR, which records all new cancer cases in the Netherlands, according to the following process. The NCR receives lists of newly diagnosed patients on a regular basis from the pathology departments. In addition, the medical records departments of the hospitals provide lists of outpatient and hospitalized cancer patients. Following this notification, the medical records of newly diagnosed patients (and tumors) are collected, and trained registrars from the cancer registry abstract the required information. Data are checked for duplicate records.

Several possible time lags in this process must be taken into account, and the duration of these time intervals may vary by cancer type: (1) interval from the time a case is diagnosed clinically in a clinician's office or inpatient facility to when the case is initially reported to the cancer registry; (2) interval from the initial report to the cancer registry to the time full reporting is completed (including complete stage information and details of initial treatment); and (3) interval from the time a case is partly or fully recorded in the cancer registry to the time the data are available in the PHARMO linkage. It is estimated that data validating diagnoses of breast cancer for this study should be available within approximately 4 months after initial reporting and that data validating diagnoses of bladder cancer should be available within approximately 6 months to 1 year after initial reporting. Analyses to address the study objectives will be conducted using only confirmed cases; since we do not anticipate that there will be a category of provisional cases in PHARMO, separate analyses using confirmed and provisional cases combined will not be necessary.

HIRD^{SM} and Medicare

Validation of provisional cases in HIRD^{SM} and Medicare data will be performed through review of medical records. Cases will be considered confirmed if there is supportive evidence of a cancer diagnosis; that is, recorded treatment or procedure codes for cancer-specific therapy (surgery, radiation therapy, chemotherapy, hormonal therapy, or other targeted or biological therapy), within the period from 1 month before to 3 months after at least two recorded diagnostic codes for female breast cancer or bladder cancer (males and females).

Patient identifiers (name, date of birth, and social security number) can be requested from CMS. If the request is approved, these can be used for further data abstraction. Additionally, each individual or institutional provider has a unique identification number that is used to identify specific providers. A sample of relevant potential cases will be identified from each cohort, and a list will be sent to separate trusted third parties for Medicare and for HIRD^{SM}. Each third party will contact the individual provider to obtain the required information from relevant medical records. For patients in the Medicare database, details from the medical record will be obtained by record abstraction. For patients in the HIRD^{SM}, redacted copies of pertinent portions of the medical record will be obtained. Structured forms for abstraction in Medicare and for guiding the copying of relevant records for HIRD^{SM} patients will be used to collect the relevant information to confirm the outcome (these forms will be provided as part of the study report). Final confirmation of cases will be conducted independently by endpoint adjudicators who will be blinded to exposure to medications and will be identified by RTI Health Solutions (RTI-HS) and HealthCore.

3.4.2 Exposure/Independent Variables of Interest

Exposure will first be classified according to treatment at the index date ("as initiated"). For a patient in the comparator group who subsequently qualifies for entry into the dapagliflozin group, person-time in the comparator group will be censored at the time the patient qualifies for entry into the dapagliflozin group and subsequent person-time will be classified as exposed to dapagliflozin. By contrast, for a patient in the dapagliflozin group, person-time will not be censored if a comparator AD is started. Therefore, the person-time exposure classification "as initiated" is equivalent to a dichotomized categorization of patients' person-time as being "ever exposed" to dapagliflozin versus "not yet exposed" to dapagliflozin.

For many known carcinogens and promoters, cancer risk increases with cumulative exposure. Therefore, in addition to the "as initiated" exposure classification already described, we will analyze the effect of dapagliflozin according to mutually exclusive categories of cumulative exposure. Since there is no variable for days of medication supplied in the prescription information in the CPRD, we will initially use as a proxy for cumulative exposure each patient's recorded number of prescriptions in the CPRD. The recorded number of dispensings will be used similarly in PHARMO data. Using the number of prescriptions or dispensings as a proxy for cumulative exposure assumes that most prescriptions were written for a standard period of use (typically 1 month) and that the prescribed daily dosage of dapagliflozin (recommended dose is 10 mg dapagliflozin once daily) does not vary substantially among prescriptions/dispensings. The validity of these assumptions will be assessed using the PHARMO Outpatient Pharmacy Database, in which the duration of use for each dispensing can be calculated by dividing the number of units dispensed by the number of units to be used per day as defined in the pharmacies. We will evaluate whether there is evidence of any trend of increasing cancer risk with increasing cumulative exposure to dapagliflozin (see further discussion in Section 4.1, Statistical Analysis Methods). In HIRD^{SM} and Medicare data, the number of days' supply, when available, will be summed to estimate cumulative exposure.

3.4.3 Other Covariates/Control Variables

The following data will be collected from the data sources for all study patients at the index date whenever available: age, sex, calendar year, geographic region, duration of lookback time, socioeconomic status, medical comorbidities, and concomitant medications. The following health service utilization data will be collected for a period of 180 days before but not including the index date: number of physician outpatient, emergency department, hospital, or specialty care visits. Duration of T2DM and measures of the degree of control of metabolic abnormalities in patients with T2DM are potentially important covariates; however, this information will not be evaluable in HIRD^{SM} and Medicare data and may not be evaluable consistently in the CPRD and PHARMO databases.

We will collect information on whether patients received prior antidiabetic therapy and/or if they "added on" or were "switched to" dapagliflozin or other ADs at the time they initiated treatment. Changes in treatment will be treated as time-dependent variables. We will evaluate these "added on," "switched to" or other ADs as potential confounders or effect modifiers. If sample size permits, the "added on" and "switched to" populations will be analyzed separately. These covariates will be evaluated as potential effect modifiers and confounders at the time the study drug (dapagliflozin or comparator) was initiated (prescription index date) rather than at the time of cohort entry; therefore, they will not be included in the propensity scores.

Use of additional ADs (other than the AD that qualified a patient for cohort entry) will be monitored during follow-up as potential confounders. To be considered exposed to an additional

AD during follow-up, as a time-varying covariate in the outcome model, at least one prescription for the medication must be recorded during follow-up (recorded as yes/no for whether a particular AD was prescribed).

The statistical analysis plan will include a complete list of covariates that can be ascertained from the data sources that will be used for this study. It is important to note that electronic health care data sources will not include information for all possible risk factors. In particular, information regarding family history and lifestyle (i.e., exercise habits, alcohol consumption, and cigarette smoking) may not be captured or may not be available to investigators. In HIRD^{SM} and Medicare data, smoking history and alcohol consumption will not be available.

Covariate values will be estimated at the index date for propensity score estimation. In addition, time-varying confounders will be considered for inclusion in Cox regression analysis. Lists of covariates that are available in each data source are provided, by cancer type, in Table 1 through Table 3.

Table 1: Breast Cancer Covariates, Information in Data Source

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Demographics and lifestyle			
Age	Yes	Yes	Yes
Overweight/obese	BMI available[a]	For 15% of the patients, available via either clinical lab data or GP	Only interventions on claims, as surrogates
Alcohol consumption	Diagnostic codes for alcohol abuse	Available via GP (only alcohol abuse)	No
Tobacco use	Current, former, or nonsmoker[a]	Available via GP	No
Medical comorbidities			
Benign mammary dysplasia	Diagnostic and pathology codes	Via pathology database	Diagnosis codes
Renal insufficiency	Diagnostic codes	Via hospital admission database	Diagnosis codes
Retinopathy	Diagnostic codes	Via hospital admission database	Diagnosis codes
Peripheral neuropathy	Diagnostic codes	Via hospital admission database	Diagnosis codes
Peripheral vascular disease	Diagnostic codes	Via hospital admission database	Diagnosis codes
Cardiovascular disease	Diagnostic codes	Via hospital admission database	Diagnosis codes
Cerebrovascular disease	Diagnostic codes	Via hospital admission database	Diagnosis codes
Hospitalizations	Flag on encounters	Via hospital admission database	Yes

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Amputations	Diagnostic codes	Via hospital admission database	Diagnosis codes
Medications			
Hormone-replacement therapy	Prescription codes	Via outpatient pharmacy	Outpatient dispensings
Baseline antidiabetic treatments	Prescription codes	Via outpatient pharmacy	Outpatient dispensings
Selective estrogen receptor modulators (raloxifene tamoxifen) (reduces risk)	Prescription codes	Via outpatient pharmacy	Outpatient dispensings
Opioids	Prescription codes	Via outpatient pharmacy	Outpatient dispensings

BMI = body mass index; CPRD = Clinical Practice Research Datalink (UK); GP = general practitioner; HIRD^{SM} = HealthCore Integrated Research Database; PHARMO = PHARMO Database Network (the Netherlands).

[a] Data on body weight and height and smoking were missing for approximately 30% of patients in one CPRD study (Gelfand et al., 2005).

Table 2: Bladder Cancer Covariates, Information in Data Sources

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Demographics and lifestyle			
Age	Yes	Yes	Yes
Sex	Yes	Yes	Yes
Overweight/obese	BMI available[a]	For 15% of the patients, available via either clinical lab data or GP	Only interventions on claims, as surrogates
Alcohol consumption	Diagnostic codes for alcohol abuse	Available via GP (only alcohol abuse)	No
Tobacco use	Current, former, or nonsmoker[a]	Available via GP	No
Medical comorbidities			
Hereditary nonpolyposis colon cancer	Diagnostic and pathology codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Urinary infections (chronic or recurring)	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Chronic or recurring urinary cystitis	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Recurrence of relevant diagnosis codes
Kidney stones	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Bladder stones	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Renal insufficiency	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Retinopathy	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Peripheral neuropathy	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Peripheral vascular disease	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Cardiovascular disease	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Cerebrovascular disease	Diagnostic codes	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Diagnosis codes
Hospitalizations and amputations	Diagnostic codes and flag on encounters	1) Outpatient pharmacy data: medication use as proxy; 2) Hospital admission data: only severe cases; 3) GP data (subcohort)	Hospital claims and diagnosis codes
Medications			
Cyclophosphamide	Prescription codes	If orally dispensed via public pharmacy data. Initial chemotherapy yes/no via NCR	Code on outpatient dispensing

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Radiation therapy in pelvic region	Not specifically recorded but may be in GP notes or diagnostic codes for complications	Initial radiation therapy via ECR	Diagnosis or procedure codes
Opioids	Prescription codes	Via outpatient pharmacy	Outpatient dispensings
Baseline antidiabetic treatments	Prescription codes	Via Outpatient Pharmacy Database	Codes on outpatient dispensings

BMI = body mass index; CPRD = Clinical Practice Research Datalink (UK); NCR = Netherlands Cancer Registry; GP = general practitioner; HIRD^{SM} = HealthCore Integrated Research Database; PHARMO = PHARMO Database Network (the Netherlands).

[a] Data on body weight and height and smoking were missing for approximately 30% of patients in one CPRD study (Gelfand et al., 2005).

Table 3: Other Cancer Covariates, Information in Data Sources

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Demographics and lifestyle			
Age	Yes	Yes	Yes
Sex	Yes	Yes	Yes
Overweight/obese	BMI available (with some “missing”)	For 15% of the patients, available via either clinical lab data or GP	Only interventions on claims, as surrogates
Alcohol consumption	Diagnostic codes for alcohol abuse	Available via GP (only alcohol abuse)	No
Tobacco use	Current, former, or nonsmoker (with some “missing”)	Available via GP	No
Medical comorbidities			
Polycystic ovarian syndrome	Diagnostic codes	Via pathology database or Hospitalization Database	Diagnosis codes
Colon polyps	Diagnostic and pathology codes	Via pathology database or Hospitalization Database	Diagnosis codes
Crohn’s disease	Diagnostic and pathology codes	Via pathology database or Hospitalization Database	Diagnosis codes
Ulcerative colitis	Diagnostic and pathology codes	Via pathology database or Hospitalization Database	Diagnosis codes

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Pancreatitis	Diagnostic codes	Via pathology database or Hospitalization Database	Diagnosis codes
Immunosuppressive diseases such as HIV/AIDS	Prescription codes	1. HIV via clinical lab; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Outpatient pharmacy data: medication use as proxy	Diagnosis codes
Renal insufficiency	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Retinopathy	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Peripheral neuropathy	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Peripheral vascular disease	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Cardiovascular disease	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Cerebrovascular disease	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Hospitalizations	Flag on encounters	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Hospital claims and diagnosis codes
Amputations	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Helicobacter pylori infection	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes

Type of Information	CPRD	PHARMO	HIRD^{SM} and Medicare
Autoimmune diseases	Diagnostic codes	1. Outpatient pharmacy data: medication use as proxy; 2. Hospital admission data: only severe cases; 3. GP data (subcohort); 4. Clinical lab	Diagnosis codes
Medications			
Hormone-replacement therapy	Prescription codes	Via outpatient pharmacy	Codes on outpatient dispensings
Unopposed estrogen therapy	Prescription codes	Via outpatient pharmacy	Codes on outpatient dispensings
Selective estrogen receptor modulators (raloxifene, tamoxifen)	Prescription codes	Via outpatient pharmacy	Codes on outpatient dispensings
Immunosuppressant such as steroids	Prescription codes	Via outpatient pharmacy	Codes on outpatient dispensings
Opioids	Prescription codes	Via outpatient pharmacy	Outpatient dispensings
Baseline antidiabetic treatments	Prescription codes	Via outpatient pharmacy	Codes on outpatient dispensings

BMI = body mass index; CPRD = Clinical Practice Research Datalink (UK); GP = general practitioner; HIRD^{SM} = HealthCore Integrated Research Database; HIV = human immunodeficiency virus; PHARMO = PHARMO Database Network (the Netherlands).

4 STATISTICAL ANALYSIS

4.1 Statistical Analysis Methods

The primary objectives of these analyses are (1) to compare, by insulin use at the index date, the incidence of breast cancer among females with T2DM who are new users of dapagliflozin with the incidence among females who are new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy and (2) among both sexes, to similarly compare the incidence of bladder cancer between these exposure groups.

Secondary objectives are to compare, between the previously defined exposure groups, the following characteristics during follow-up, by insulin use at the index date: (1) the frequency of several measures of health care utilization (including outpatient visit frequencies and use of breast and bladder cancer screening and diagnostic tests); (2) baseline patient characteristics between the groups; (3) the composite incidence of selected invasive cancers (prostate, colon/rectum, lung, stomach, NHL, and melanoma of skin) among males in these groups; and (4) the composite incidence of selected invasive cancers (colon/rectum, lung, corpus uteri, ovary, stomach, NHL, and melanoma of skin) among females in these groups.

Specifics of variable definitions will be described in the statistical analysis plan (SAP), to be developed after finalization of the protocol.

Three analyses will be done with respect to included follow-up time and cancer cases. One analysis will include all follow-up time and all cancer cases. This analysis should provide the most sensitivity to detect surveillance bias of prevalent cancers that may be diagnosed in relation to starting a study AD drug. Another analysis will restrict the follow-up time (and cancer cases) to that accrued more than 6 months after the index date. A third analysis will restrict the follow-up time (and cancer cases) to that accrued more than 1 year after the index date. The second and third analyses should show progressively reduced effects of surveillance (detection of prevalent cases) related to starting a study AD drug.

All conversion of the original data to analysis variables will be performed using SAS software, version 9.2 or higher (Cary, NC: SAS Institute Inc.; 2008).[3] Data management will be carried out in accordance with RTI-HS, PHARMO, and HIRD^{SM} standard operating procedures. Routine procedures include checking electronic files, maintaining security and data confidentiality, following the SAP, and performing quality-control checks of all programs. Researchers at RTI-HS will be responsible for analyzing data from the CPRD and Medicare, and researchers at PHARMO and the HIRD^{SM} will analyze data from these organizations. Data extraction programming for creating the study population from the HIRD^{SM} and creating the analytic file will be performed in accordance with HealthCore Programming Standards. The HealthCore Programming Standards are a set of documents describing data extraction and data development methods that are referenced in HealthCore standard operating procedures.

Given the published findings suggesting an association between pioglitazone and bladder cancer (Lewis et al., 2011), all analyses relating to bladder cancer in the present study will be conducted with and without users of pioglitazone in the dapagliflozin-exposed and comparator cohorts. If sufficient data are available, a subgroup analysis in concomitant users of dapagliflozin and pioglitazone may be carried out to determine how much, if any, increase over additivity of effects there is among patients exposed to both agents.

4.1.1 Propensity Score Approach

Demographic, medical, and clinical factors that may be associated with the decision to begin therapy with a particular AD may also be associated with the outcome. However, the number of outcomes will likely be small, limiting the number of variables that could be included in a regression model that predicts these outcomes (Cepeda et al., 2003). To address this difficulty, we will summarize the set of confounding variables into a single propensity score, based on knowledge of potential confounding variables associated with cancer risk. The propensity score is the predicted probability of being assigned to a particular treatment conditional on a set of observed covariates (Braitman and Rosenbaum, 2002; D’Agostino, 1998; Perkins et al., 2000).

Within each data source, propensity scores will be estimated by conducting logistic regression modeling and incorporating measured potential predictors of therapy as independent variables.

[3] Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

The outcome variable in the propensity score model is exposure status (dapagliflozin-exposed vs. comparator cohort). Selection of variables to adjust and include in the propensity score modeling will be factors that are associated with a reported increase or decrease in cancer risk. Variables to be considered for propensity score derivation, which include demographic, clinical (medical comorbidities and concomitant medications), and health care utilization variables present at or before the index date, are listed in Table 1, Table 2, and Table 3; they will be assessed if available in the data source. Indicator variables for the duration of lookback time and timing of information on key covariates may also be included. As noted in the above tables, in the US HIRD^{SM} and Medicare databases, some variables such obesity, smoking, and alcohol consumption will not be available. However, it should be noted that some clinical diagnosis variables that would be useful in estimating propensity scores in a study of cancer outcomes will not be available in the linked PHARMO outpatient pharmacy, hospitalization, clinical laboratory, pathology, and cancer registry data. To the extent possible, drug prescription proxies for diagnoses will be identified.

In the CPRD, if it is not feasible to obtain data for all eligible comparators, a subsample of AD users will be identified prior to propensity score estimation. The subsample will be identified by frequency matching eligible comparator patients to dapagliflozin patients in at least a 6:1 ratio by 5-year age groups, sex, geographic region, and calendar year of the index date. Similarly, in Medicare, if it is not possible to obtain data from all available comparators, a subsample of AD users will be randomly sampled in at least a 15:1 ratio to dapagliflozin patients using the same matching criteria. For the HIRD^{SM} and PHARMO, all patients' data will be used to estimate propensity score models before stratification. Operationally, propensity score strata will then be formed separately for each data source. For all analyses, we will exclude patients who have estimated propensity scores outside the range that is common to both dapagliflozin-exposed and comparator cohorts. This process is known as "trimming." Trimming occurs at both ends of the propensity score scale. At the lower end, we will exclude all patients, exposed or unexposed, who have a propensity score below the 2.5 percentile value of the distribution of scores among the exposed group. At the upper end, we will exclude all patients, exposed and unexposed, with scores greater than the 97.5 percentile of scores among the comparators.

Within each set of propensity scores, after trimming, the data will be stratified into deciles of propensity scores based on the distribution among new users of dapagliflozin. Within each of these 10 propensity score–based strata, we will investigate the extent to which covariates are balanced between the two treatment groups by visual inspection of the distribution of covariates and use of the absolute standardized difference to assess the balance of measured baseline covariates between the dapagliflozin group and the comparator AD group before and after propensity stratification (Austin, 2009). Any imbalance will be addressed by either revising the propensity score model or by making adjustments in the final outcome model (Braitman and Rosenbaum, 2002; Perkins et al., 2000). We will report the number of patients trimmed from the analysis because of nonoverlap of propensity scores. If using deciles to create strata results in strata that are too small, it may be necessary to combine adjacent deciles within a given year of the index date.

4.1.2 Primary Objective #1 – Calculation and Comparison of Female Breast Cancer Incidence Rates

Analyses will initially be conducted separately within each data source. The incidence of breast cancer cases among females after the index date will be estimated in the dapagliflozin-exposed and comparator cohorts. The following incidences and comparisons will be generated:

- Crude incidence and incidence rate ratio (IRR), estimated by insulin use at the index date, among the dapagliflozin-exposed group versus those unexposed to dapagliflozin
- Propensity score–adjusted incidence and IRR, estimated by insulin use at the index date, among the dapagliflozin-exposed group versus those unexposed to dapagliflozin
- Incidence and trend in incidence according to cumulative exposure to dapagliflozin, estimated by insulin use at the index date
- Cumulative incidence function graphs for breast cancer in the two groups

The adjusted IRRs will be the primary endpoint. In CPRD and Medicare data, crude IRRs will facilitate comparison with the adjusted IRRs to provide an indication of the degree of confounding.

The number of new cases of breast cancer during follow-up will be determined within each data source. Person-time for each patient will be determined as the time between the index date and the date of first diagnosis of breast cancer, the last day of available follow-up in the data source, death, end of study, end of the risk window for the index AD, or the first diagnosis of any other invasive cancer, whichever occurs first. The total person-time of observation among individuals at risk will then be calculated. The incidence rate of breast cancer will be estimated by insulin use at the index date in each cohort. Crude and adjusted rates will be calculated as the number of new cases of disease during the observation period divided by the total person-time of observation among individuals at risk. Incidence rates will be reported as point estimates (cases per 10,000 person-years) and 95% CIs. Adjusted incidence rates will be adjusted at least by propensity score and will be derived separately by data source.

Cumulative dose of dapagliflozin use will be measured as described previously. The incidence rate of breast cancer will be calculated within categories of cumulative dose as the number of new cases of disease during the follow-up period in a given category of cumulative dose divided by the total person-time observed in this cumulative dose category. We will investigate whether there is any trend of increasing cancer risk with increasing cumulative dose of dapagliflozin. In this analysis of breast cancer rate by cumulative dose of dapagliflozin, patients in the comparator cohort will be analyzed in a dose category of zero. Because selection bias for dapagliflozin treatment or residual confounding may apply preferentially to the zero dose category more than other dose categories, we will also explore whether omitting the zero dose category yields a substantially different trend analysis. Results will be reported as a point estimate (cases per 10,000 person-years) and 95% CI. Separate analyses will be conducted for confirmed cases and for confirmed plus provisional cases combined.

4.1.3 Primary Objective #2 – Calculation and Comparison of Bladder Cancer Incidence Rates

Bladder cancer overall and sex-specific incidence will be analyzed similarly to breast cancer incidence, as described in the preceding section. If a sufficient number of pioglitazone-exposed patients are included in the study, we will explore whether the incidence of bladder cancer in those exposed to both pioglitazone and dapagliflozin is estimated to be higher than the expected additive effect of the two exposures independently.

4.1.4 Secondary Objective #1 – Frequency and Comparison of Health Care Utilization During Follow-up

Measures of health care utilization during follow-up will be compared between new users of dapagliflozin and new users of comparator ADs within the CPRD, the HIRD^{SM}, and Medicare data and will be stratified by insulin use at the index date. The use and frequency of diagnosis-specific tests that may lead to a diagnosis of breast or bladder cancer, including cystoscopy, mammography, and breast biopsy, will be summarized for the dapagliflozin-exposed and comparator groups to describe the medical surveillance intensity and to determine whether the pattern of tests that might lead to a study outcome diagnosis differ between exposure groups. Categorical utilization variables will be summarized by frequencies and proportions, and continuous variables will be summarized by means and standard deviations or medians and interquartile ranges.

When appropriate, we will evaluate health care utilization that is associated with a diagnosis of cancer using proportional hazards regression. We will examine the strength of the association (expressed as unadjusted hazard ratios with 95% CIs) of dapagliflozin use between patients with and without each factor of interest. Variables that are associated with a change in the estimate of the effect of dapagliflozin use by 10% or more in relative risk (crude analysis vs. analysis stratified on variable of interest) will be included in multivariable proportional hazards regression models to determine independent predictors of cancer diagnosis (estimated as adjusted hazard ratios with 95% CIs).

4.1.5 Secondary Objective #2 – Frequency and Comparison of Baseline Patient Characteristics

We will tabulate and compare patient characteristics at the index date stratified by baseline insulin use between patients with T2DM who are new users of dapagliflozin and those who are new users of ADs in classes other than SGLT2 inhibitors, insulin monotherapy, metformin monotherapy, or sulfonylurea monotherapy.

Descriptive statistics will be generated within each data source to compare baseline characteristics between dapagliflozin users and comparator AD users, separately within categories of baseline insulin. Categorical variables will be summarized by frequencies and proportions, and continuous variables will be summarized by means and standard deviations or medians and interquartile ranges. The following variables will be characterized, where available:

- Age stratified by sex

- Duration of history in data source before the index date
- Summary of all AD medications used at the index date (other than dapagliflozin)
- Switch versus add-on initiation
- Smoking history
- Family history of cancer
- History of chronic obstructive pulmonary disease (COPD) or asthma
- History of alcoholism or alcoholic liver disease
- History of human immunodeficiency virus (HIV)/AIDS

Results of the descriptive analyses will be used to inform the stratification of subsequent analyses.

In the CPRD, outpatient diagnoses and prescriptions will be used to ascertain the medical history. In the PHARMO Database Network, hospital discharge diagnoses for all patients and outpatient diagnoses for a small subcohort will be available; however, medical conditions that are treated, such as hypertension, can be identified from pharmacy dispensings. In HIRD^{SM} and Medicare data, outpatient diagnoses and prescription and inpatient discharge diagnoses will be used to ascertain medical history. Selection of variables to adjust and include in the propensity score modeling will be factors that are associated with a reported increase or decrease in cancer risk based on available literature. After cases are ascertained, we will assess whether factors included in the propensity score model are predictors of the primary outcomes within the study population.

4.1.6 Secondary Objectives #3 and #4 – Comparison of Composite Cancer Incidence Rates for Selected Cancers Among Males and Females Separately

The composite incidence of the selected invasive cancers will be determined within each data source for males and females separately. These analyses will be similar to those for the primary objectives (incidence of female breast cancer and of bladder cancer in both sexes). For the analysis of the composite endpoints, a patient will be followed until the first occurrence of any of the cancers that the endpoint comprises. Therefore, person-time accumulation will be calculated separately for these analyses compared with those for the primary study objectives. The incidence rates of the composite cancer endpoint will be determined in each exposure group, stratified by propensity score, as the number of new cases of any of the selected cancers during the follow-up period divided by the total person-time observed after the index date, estimated by insulin use at the index date. Each result will be reported as a point estimate (cases per 10,000 person-years) and 95% CI.

4.1.7 Evaluation of Time-Varying Factors

Time-dependent variables such as change in the severity of diabetes and changes in the intensification of diabetes treatment (i.e., based on the number of antidiabetic drugs of different classes that are prescribed simultaneously), that could represent possible confounders and effect modifiers will be assessed and classified during follow-up time. The degree to which we can pursue analyses of these factors is contingent on the number of outcome events. If any of these factors is considered an important possible confounder or effect modifier, adjustment for the factor will be performed via time-dependent Cox proportional hazards regression, implemented by categories of insulin use at the index date and including propensity score deciles at the index date, the categorical time-dependent indicator (i.e., diabetes severity and/or change in stage of diabetes treatment), and calendar year.

4.1.8 Imputation of Missing Values

We expect that relatively few key variables will have notable missing values, with the possible exception of lifestyle variables. The pharmaceutical exposures and comorbidities are expected to be based on outpatient prescriptions and to be complete. If there are considerable missing data for lifestyle covariates, multiple imputation will be used to fill in missing values for the propensity score creation and multivariable analyses. The decision to use multiple imputation will depend on the strength of the association between the variable and treatment and the extent of missing data. Based on information from the observations with nonmissing values, we will impute five simulated versions of the dataset. The imputed datasets will be used for creation of propensity scores and in the multivariable analyses, with the results being combined appropriately to generate final point estimates and CIs. In theory, this should give point estimates with equal or less bias than those that would be obtained if we had limited the sample to those with complete data, and it should give greater precision because of the larger number of patients that will be included using this method as opposed to restricting the analysis to observations with complete data. The specific approach will be detailed in the statistical analysis plan.

We have selected the multiple imputation approach because existing methods for imputation penalize the standard errors when imputing data and multiple imputation allows for better bias correction than most alternatives, including the complete-case approach, for many, although not all, applications. The complete-case approach can be very costly of information in a body of high-dimensional data, since the proportion of complete cases will decline with the increase in the number of variables.

4.1.9 Sensitivity Analyses

The effect of some carcinogens and promoters on cancer risk decreases after exposure is discontinued. Therefore, if a sufficient number of patients accumulate a relatively high cumulative exposure to dapagliflozin and are observed to experience an increased cancer risk, we will explore whether such risk decreases with time since discontinuation (conditional on the category of cumulative dose received). Since the analysis is stratified by cumulative exposure, its utility depends on having a sufficient number of patients with substantial cumulative exposure to dapagliflozin who subsequently discontinued its use. If there are few such patients, strata with high cumulative exposure and varying times since discontinuation will be sparsely populated and this analysis will add little information to the cumulative exposure analysis.

If it is suspected that there is residual confounding in one or both data sources (for example, due to lack of information on one or more confounding variables), an approach that can be used to reduce the amount of such confounding is external adjustment of the estimate from analysis of each data source study with residual confounding (Lash and Fink, 2003). We will assess the effect of unmeasured confounders, one at a time, on the association between dapagliflozin use and the primary cancer outcomes by assuming a plausible range of values for the prevalences of each of the unmeasured confounders among the dapagliflozin group and the comparator group and risk ratio for the association between each of the unmeasured confounders and the outcome of interest (Chapter 5 in Lash et al., 2009). For example, female reproductive risk factors such as age at menarche or whether a woman breast fed can be risk factors for breast cancer and are likely to be unmeasured. Based on the available literature, we can assume a reasonable range of prevalence values for a given unmeasured confounder and a specific relative risk for the association of the risk factor and the outcome of interest to give a range of modified values for the associations between exposure and outcomes observed. However, at this time there is no reason to expect differential distribution of reproductive risk factors between dapagliflozin and other AD users among the study population.

Sensitivity analyses without stratification on categories of insulin use will also be conducted. Other sensitivity analyses will explore the potential for surveillance bias by evaluating the risk of cancer in relation to time since first exposure to the study drugs and by a lag time analysis to assess the temporal relation of risk to exposure that restricts analysis to different windows of follow-up time after the index date. To explore the effect of potential differences in the risk of death between the dapagliflozin cohort and the comparator AD cohort, estimation of all-cause mortality rates, accounting for person-years at risk in each cohort, will be performed where feasible. The mortality rate ratio and 95% confidence intervals will be calculated. Cause-specific death rates will not be calculated because data on underlying cause of death are not recorded in all data sources. An additional sensitivity analysis will be conducted that does not censor follow-up at the occurrence of any of the cancer endpoints. Patients will continue to be eligible to experience a study outcome, even if they experienced a different type of cancer earlier in follow-up. In this approach, all breast and bladder cancers will be identified. However, limitations in the ability to distinguish between primary and secondary cancers should be taken into account.

4.1.10 Pooled Analyses

If the results of the study across all four data sources are similar for at least one of the primary outcomes (i.e., plausibly differing only from sampling variability), techniques will be used to pool the data from the different data sources. The pooled analysis will be designed to estimate the effect of the exposure while controlling for confounding using the data stratified on propensity scores. The data source will be retained as a stratification variable, so the effect within each data source can be estimated.

Mantel-Haenszel techniques will be used to pool the data from each data source and calculate overall adjusted incidence rate ratios. This analysis is designed to estimate the effect of the exposure while controlling for confounding by using data source–specific propensity score stratification.

If it is suspected that there is residual confounding in any of the data sources (for example, due to lack of information on one or more confounding variables), external adjustment can be used to reduce the amount of such confounding (see Section 4.1.9) (Lash and Fink, 2003).

4.1.11 Power/Sample Size

The observed study size will depend upon the market uptake of dapagliflozin. Currently, we estimate that in the CPRD at the end of 10 years, there will be approximately 9,500 person-years of follow-up available among all new users of dapagliflozin. These estimates are based on the following assumptions: average number of patients aged 40 years and older with a newly prescribed AD with at least 180 days of prior enrollment was 10,150 per year in the CPRD (CPRD data as of 31 December 2011); the proportion of new users starting dapagliflozin among patients who meet these inclusion criteria will be 1%, 2%, 3%, 4%, and 5% during the first 5 years of the study and 5% for each subsequent year; and the annual loss to follow-up will be 5%. Using similar assumptions, we estimate approximately 5,800 person-years of follow-up available among new users of dapagliflozin in PHARMO databases after 10 years. Assuming approximately 20% of new dapagliflozin users will be on insulin at the index date (Hall et al., 2012), the CPRD will contribute approximately 7,600 person-years of follow-up from those not on insulin at the index date and 1,900 from those on insulin at the index date; and PHARMO will contribute 4,640 and 1,160 person-years in these categories, respectively. If women contribute approximately half of these person-years of follow-up, after 10 years the CPRD will provide a total of approximately 3,800 female person-years of follow-up among women who were not on insulin at their index date.

In the US, we estimate that there will be approximately 835,000 person-years of follow-up available among all new users of dapagliflozin (138,000 person-years in the HIRD^{SM} and 697,000 person-years in Medicare data) over 9 years. This exposure would include approximately 668,000 person-years among those not on insulin at the index date and 167,000 person-years among those on insulin at the index date. If females contribute half of the person-time, we expect 55,000 exposed person-years for females not on insulin in the HIRD^{SM} and 279,000 person-years in Medicare data. These estimates are based on the following assumptions: (1) 37.7% of oral AD users are aged 65 years or older (Boccuzzi et al., 2001); (2) 80% of oral AD users aged < 65 years are aged 40-64 years; (3) of the US population aged 65 or more years, 34.15% are covered by Medicare Part D in a non-managed care program, so their data will be available for research; (4) the HIRD^{SM} covers 6% of the US population aged 40-64 years; (5) each new user will contribute 1.5 years of person-time in the HIRD^{SM}, and there will be no turnover in Medicare after the initial 180-day assessment for comorbidities etc.; (6) 80% of patients in the HIRD^{SM} and 92% of the patients in Medicare will have at least 180 days of enrollment eligibility prior to starting the drug; and (7) approximately 20% of new dapagliflozin users will be on insulin at the index date (Hall et al., 2012).

Therefore, we estimate there will be a total of 850,000 person-years of follow-up among new users of dapagliflozin across all four data sources, including 425,000 person-years of exposed follow-up among all new female users and 340,000 person-years of exposed follow-up among females not on insulin.

To provide precision estimates in relation to the projected study size based on the breast and bladder cancer outcomes, we first estimated the background incidence rates using data from the UK and the Netherlands as reported in GLOBOCAN 2008 (http://globocan.iarc.fr/, GLOBOCAN online analysis tool; accessed 20 July 2012) and from the US as reported by the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute. The expected incidence of breast cancer among women with diabetes aged 40 years and older was estimated as the background rate for this age group in the general population multiplied by 1.2, which is the relative risk reported from a meta-analysis of 20 epidemiologic studies of breast cancer risk among women with diabetes compared with women without diabetes (Larsson et al., 2007). The resulting estimated incidence of breast cancer is 34 per 10,000 female person-years in Europe, 49.5 per 10,000 female person-years in the US among those aged 65 and older, and 9.4 per 10,000 female person-years in the US among those aged less than 65 years. If this is the rate in the new users of dapagliflozin not on insulin at the index date, then we would expect to observe approximately 1,450 events in females across all data sources if the 340,000 person-years we anticipate are accrued among this group of exposed patients not on insulin. Table 4 shows the expected number of breast cancer cases among all female cohort members by study data source.

Table 4: Estimated Number of Cancers by Study Data Source Among All Cohort Members

	HIRD^{SM}	**Medicare**	**CPRD**	**PHARMO**	**Total**
Total sample cohort (person-years)[a]	689,000	3,484,000	47,500	29,000	4,250,000
Females (person-years)	345,000	1,742,000	23,750	14,500	2,125,000
Female Breast Cancer					
Rate of breast cancer (per 10,000 person-years)	9.4	49.5	34	34	
Estimated number of events	325	8,630	81	49	9,085
Bladder Cancer					
Rate of bladder cancer (per 10,000 person-years)	0.7	17.3	4.6	4.6	
Estimated number of events	50	6,010	22	13	6,095

CPRD = Clinical Practice Research Datalink; HIRD^{SM} = HealthCore Integrated Research Database; PHARMO = PHARMO Database Network, the Netherlands.

[a] Assumes 34.15% coverage rate for Medicare Part D, and 6% population coverage of HIRD^{SM}, 1.5 year of follow-up in the HIRD^{SM}, and 80% (HIRD^{SM}) and 92% (Medicare) of patients with 180 days enrollment prior to first dapagliflozin use. Estimated person-years represent a conservative scenario based on a ratio of 4 AD comparator initiators for each individual dapagliflozin initiator.

To estimate the anticipated magnitude of the upper confidence interval, we calculated a weighted average of the three expected background incidence rates (9.4 per 10,000; 49.5 per 10,000; and 34 per 10,000), based on the age distributions and the distribution of person-years in each data source. We estimated the precision of the study under various scenarios using this weighted incidence rate.

Table 5 provides for breast cancer the probability that the upper confidence limit around the observed IRR will be less than specified IRRs assuming that the true IRR is 1.0. For example, a study size of 200,000 person-years of dapagliflozin follow-up among female new users of dapagliflozin not on insulin at the index date will provide a 71% probability that the upper 95% confidence limit of the IRR will be less than 1.1.

Table 5: Probability That Upper 95% Confidence Limit of IRR for Female Breast Cancer is Below Specified Value, Assuming IRR in Population = 1.0

Female Person-years of Dapagliflozin Exposure	Upper 95% CL of Incidence Rate Ratio for Dapagliflozin Versus Other Antidiabetic Drugs				
	1.05	**1.1**	**1.15**	**1.2**	**1.3**
50,000	0.09	0.24	0.45	0.67	0.93
100,000	0.15	0.42	0.74	0.92	1.00
200,000	0.25	0.71	0.96	1.00	1.00
400,000	0.44	0.94	1.00	1.00	1.00
700,000	0.67	1.00	1.00	1.00	1.00

CL = confidence limit; IRR = incident rate ratio.

Note: Assuming 42.8 per 10,000 person-years is the rate of breast cancer among female patients not exposed to dapagliflozin, a 1:4 dapagliflozin:comparator person-year ratio, and population IRR = 1.0. This table was calculated using Episheet (Rothman, 2012).

Similarly, the expected incidence of bladder cancer among patients with diabetes (both sexes) aged 40 years and older was estimated as the background rate for this age group multiplied by 1.4, which is the relative risk associated with diabetes reported from meta-analysis of seven case-control and three cohort studies of bladder cancer risk (Larsson et al., 2006). The resulting estimated incidence of bladder cancer is 4.6 per 10,000 person-years in Europe, 17.3 per 10,000 person-years in the US in patients aged 65 years or older, and 0.7 per 10,000 in patients aged younger than 65 years. Table 4 shows the expected number of bladder cancer cases among all cohort members by study data source. We calculated a weighted average of the three expected background incidence rates, based on the age distributions and the distribution of person-years in each data source. We estimated the precision of the study under various scenarios using this weighted incidence rate.

Table 6 provides for bladder cancer the probability that the upper confidence limit around the observed IRR will be less than specified IRRs, assuming the true IRR is 1.0. For example, a study size of 400,000 person-years of dapagliflozin follow-up among new users of dapagliflozin not on insulin at the index date will provide an 97% probability that the upper 95% confidence limit of the IRR will be less than 1.2.

Table 6: Probability That Upper 95% Confidence Limit of Observed IRR for Bladder Cancer is Below Specified Value, Assuming IRR in Population = 1.0

Person-years of Exposure	Upper 95% CL of Incidence Rate Ratio for Dapagliflozin Versus Other Antidiabetic Drugs				
	1.1	1.2	1.3	1.4	1.5
50,000	0.11	0.28	0.51	0.72	0.87
100,000	0.17	0.50	0.80	0.95	0.99
400,000	0.53	0.97	1.00	1.00	1.00
800,000	0.83	1.00	1.00	1.00	1.00
900,000	0.87	1.00	1.00	1.00	1.00

CL = confidence limit; IRR = incident rate ratio.

Note: Assuming a weighted average of 14.4 per 10,000 person-years is the rate of bladder cancer among patients not exposed to dapagliflozin, a 1:4 dapagliflozin:comparator person-year ratio, and population IRR = 1.0. This table was calculated using Episheet (Rothman, 2012).

4.2 Milestones

Descriptive and when appropriate comparative analyses will be conducted every 2 years and at the end of the study, which is estimated to be after dapagliflozin has been on the market for 10 years. The study timelines will be aligned with the launch of dapagliflozin in the United States, which was January 2014. Descriptive analyses will be conducted to assess the characteristics of patients prescribed dapagliflozin and comparators. This information will be used to construct propensity scores and to inform about potentially important covariates. This information may also be used to refine our comparator population. On a biannual basis, a summary of study progress will be provided to regulatory authorities. Summaries will contain the amount of exposure in number of patients and patient-years, numbers of patients with events, and event rates for dapagliflozin and comparator arms. Propensity score adjustments will also be performed for each interim analysis if sample sizes permit. The hazard ratio between treatment arms will be presented with nominal 95% confidence intervals. It is likely that the number of eligible patients and the number of events will be low during the first 24 to 48 months of study accrual. The interim comparative analysis will be performed if at least two outcome events are observed in the entire study cohort (dapagliflozin or comparator AD cohort combined).

Table 7 includes details on the anticipated timing for the data cuts. Study reports of the analyzed data will be submitted to the health authorities approximately 12 months after the data cuts.

Descriptive analyses will be conducted to assess the characteristics of patients prescribed dapagliflozin and comparators. This information will be used to construct propensity scores and to inform about potentially important covariates. This information may also be used to refine our comparator population. Every 2 years, a summary of study progress will be provided to regulatory authorities. Summaries will contain the amount of exposure in number of patients and patient-years, numbers of patients with events, and event rates for dapagliflozin and comparator arms. Propensity score adjustments will also be performed for each interim analysis if sample sizes permit. The hazard ratio between treatment arms will be presented with nominal 95% confidence intervals. It is likely that the number of eligible patients and the number of events will be low during the first 24 to 48 months of study accrual.

Table 7: Milestones for Cancer Outcomes

Report	Data Cut Time After Dapagliflozin is Available to Patients in the US (Anticipated Month/Year)
Interim descriptive analysis	24 months (January 2016) (includes only CPRD, PHARMO, and HIRD^{SM})
Interim comparative analyses	48 months - (January 2018) 72 months - (January 2020) 96 months - (January 2022) (includes all data sources)
Final analysis	120 months - (January 2024) (includes all data sources)

CPRD = Clinical Practice Research Datalink; HIRD^{SM} = HealthCore Integrated Research Database; PHARMO = PHARMO Database Network, the Netherlands; US = United States.

[a] Due to data source lags, which are typically 4-6 months in the HIRD^{SM} and CPRD and 2 years in Medicare, the 24-month report will likely include data from the first 18 months of dapagliflozin use in the HIRD^{SM} and CPRD, the 48-month report will include data through the first 42 months of dapagliflozin use in the HIRD^{SM} and CPRD and 18 months of dapagliflozin use in Medicare, and the 120-month report will include data through the first 114 months of dapagliflozin use in the CPRD and HIRD^{SM} and 90 months of dapagliflozin use in Medicare.

To ensure a robust pharmacovigilance plan, we have proposed several pharmacovigilance activities for cancer, including a large cardiovascular outcome trial (CVOT). The CVOT will randomize 17,150 patients with type 2 diabetes with the primary objectives to examine safety and benefit with respect to cardiovascular death, myocardial infarction, and nonhemorrhagic stroke. Additionally, the safety objectives in the CVOT include an assessment of malignancies including bladder cancer. The CVOT is designed to provide ongoing monitoring of cancer with event-driven interim analyses, an independent blinded adjudication committee, a data monitoring committee unblinded assessment, and design elements to control for potential detection bias. For assessment of bladder cancer, the interim monitoring plan has a defined statistical criterion (a Pocock alpha spending plan for an overall two-sided significance level of 0.10 with four interim summaries), which if met at any interim, would lead to interactions with regulatory authorities. Because of the robust nature of the CVOT, we view it as the best source to evaluate dapagliflozin exposure and bladder cancers. We are proposing to conduct this pharmacoepidemiology study as a complementary pharmacovigilance measure to the CVOT.

5 STUDY LIMITATIONS/STRENGTHS

5.1 Confounding

All potential confounding variables for which data are available will be controlled to the extent possible, through the design and through the use of propensity scores. Differences in practice and differences in the availability of some data across the data sources will affect development of the propensity scores, which will be allowed to differ among the data sources. Electronic health care data sources do not include information for all possible confounders. Specifically, information regarding genetic risk factors (e.g., *BRCA1* and *BRCA2*), family history, and lifestyle (e.g., exercise habits, alcohol consumption, and cigarette smoking) are captured to a great extent in the electronic medical records comprising the CPRD but, except for the subcohort of patients for which PHARMO GP data are available, data on these risk factors are not available in the PHARMO Database Network or the HIRD^{SM} and Medicare data that will be used for this study.

As with any database study, identification of medical events is limited to data that are captured as part of the medical record and other linked sources in which data are not collected primarily for research purposes and will rely on appropriate diagnostic codes to detect events. Cancer cases can be validated in the CPRD by review of electronic codes in medical records, questionnaires to GPs, or linkage to cancer registry data. Cancer cases in the PHARMO linked cancer registry are already validated according to standard procedures. Ascertainment of cancer outcomes in claims databases such as the HIRD^{SM} and Medicare databases is challenging due to the nature of the databases, the lack of clinical precision on the diagnostic coding system used (ICD-9-CM), and the lack of information on tumor histology or confirmed pathological data. Validation of cancer cases will therefore require review of medical records for information on cancer treatments and procedures to support the validity of the cancer diagnosis.

Requiring a minimum history of only 180 days before the index date will limit our ability to control for duration of diabetes and other chronic conditions. However, this limitation has to be balanced with the ability to generate statistically meaningful numbers of exposed patients to test for associations. Increasing the length of the minimum duration of required history would limit study size further.

Multivariable analyses cannot eliminate residual confounding from unmeasured factors, as is always true for observational studies. Propensity score stratification can achieve a high degree of balance between comparison groups on the presence or absence of dozens of variables, but it may leave unbalanced the unmeasured and unknown characteristics and confounders. For example, dapagliflozin-exposed patients could experience proportionally more bladder symptoms (e.g., increased volume excretion, increased frequency of urination, urinary tract infections) that could lead to more clinical work-up but would not necessarily be reported as symptoms or clinical signs in the data sources. If this occurs, then we may see a spurious association between dapagliflozin exposure and bladder cancer. Thus, there is the possibility that the results remain affected by unmeasured confounders. However, such a confounder would have to be moderately prevalent, strongly associated with exposure to dapagliflozin, and strongly predictive of the outcome to affect the results of this study. To assess the effect of unmeasured confounders on the association between dapagliflozin use and breast and bladder cancer, we will conduct sensitivity analyses to estimate the degree of possible bias that might be present assuming a plausible range of values for such potential confounders.

Confounding by indication (or channeling bias) is a common bias in observational pharmacoepidemiology studies whereby the indication for therapy may be associated with both treatment and outcome. Since patients who receive a particular drug therapy typically have more severe disease or a perceived higher risk (due to self-selection or physician preference) than patients not on the medication, selection of treatment can be confounded with clinical and nonclinical patient factors that may be related to outcomes of interest. New medications may be prescribed differentially to healthier patients, whom physicians believe could tolerate a product with a lesser-known safety profile, or to patients who have more severe disease, have failed previous treatment regimens, or have contraindications to other drugs (e.g., thiazolidinediones are not recommended for use in patients with heart failure). New medications may also be prescribed differentially by physicians who are “early adopters” of new technologies. As much as possible, such considerations are taken into account by the propensity score, but some aspects may remain unmeasured and could result in residual confounding. Specifically, dapagliflozin could be preferentially prescribed to patients with more severe diabetes or those who have failed other therapies. Dapagliflozin could also be preferentially prescribed to patients with fewer risk factors for breast and bladder cancer. These channeling patterns could bias the hazard ratio toward or away from the null.

5.2 Detection Bias

Detection bias is characterized by systematic differences between comparison groups in how outcomes are ascertained, diagnosed, or verified. It is a potential artifact in epidemiologic data caused by the use of a particular diagnostic technique or type of equipment or through enhanced medical surveillance. For example, cancer rates in this proposed study may vary between the treatment groups not because of an actual difference in the incidence of the disease but because of differences in the frequency of medical surveillance and cancer diagnostic procedures. Such detection bias is particularly possible when the cancer under study progresses slowly or can be present for a long time without causing symptoms that would prompt medical attention. This potential problem could be magnified if the labeling for use of dapagliflozin calls attention to an elevated risk of breast or bladder cancer that was observed in clinical trials. At the end of this study, additional analyses to better characterize detection bias may be executed.

5.3 Other Sources of Bias

Misclassification bias can result if study patients are not categorized correctly with regard to exposure or outcome. We expect little misclassification with respect to exposure, since this will be determined from prescribing/dispensing records. However, actual adherence to dapagliflozin or other ADs cannot be confirmed. Further, misclassification as to whether the patient is a new user could exist if providers supplied samples of dapagliflozin or comparator ADs for varying duration to patients, at no cost, and with no record in the data source. This will vary by country and data source and could result in different results in the different data sources.

Classification of type 2 versus type 1 diabetes mellitus may also be a source if misclassification. Potential patients with evidence of type 1 diabetes mellitus (T1DM) are to be excluded. However, with the repeated health care that individuals with T1DM or T2DM require, we anticipate that classification of diabetes type will be obvious from the relative frequency of the use of these two diagnoses in individual patients.

Misclassification of the outcome will be minimized by using medical records and cancer registry data, to the extent available, to confirm clinical diagnoses of cancer. Lack of information before the minimum 180 days of history before the index date could also be a source of bias because previous diagnoses of cancer could be missed. However, additional information obtained during the 180 days before the index date will mitigate this risk. Slightly underestimated rates in the female composite cancer endpoint may occur because women not at risk of ovarian and uterine cancer rates (e.g., women with oophorectomy or hysterectomy performed) were not removed from the female cohort.

Although we acknowledge that the use of a composite endpoint using multiple cancers may mask a potential elevated risk of one cancer if there is a protective effect for another cancer or if the magnitude of the effect on one cancer is not great enough to be apparent, the intent of these analyses is to identify a potential signal and if a signal is identified, further investigation into the individual cancers may be explored.

5.4 Study Size

In any electronic data resource, a proportion of patients on antidiabetic therapy might not meet the requirement for the full, minimum 180-day period of history before the index date. This is less a problem in the UK and the Netherlands than it is in the United States, where antidiabetic therapy could be started soon after joining a health plan for people with established T2DM. Also, patients whose characteristics are such that their propensity scores fall into the trimmed tails of the propensity score distribution will not be analyzed (although their number will be reported). Although this is a minor limitation with respect to study size, it is a strength with respect to the balancing effect of stratifying by propensity score decile. These limitations may modestly decrease the available study size, but will increase the validity of the comparison.

Certain subgroup analyses may involve small numbers, which will make results imprecise. The ability to meet the sample size projections depends upon the uptake of dapagliflozin. It is currently projected that at the end of 10 years, the combined data sources will have approximately 850,000 person-years of exposure to dapagliflozin, approximately 680,000 of those not on insulin at the index date and 170,000 of those on insulin at the index date.

5.5 Generalizability

The study results will be generalizable to patients with T2DM in the UK, Netherlands, or US who would have met the inclusion and exclusion criteria. Using medical record and cancer registry data from primary care practices in the UK and the Netherlands (instead of limiting the study to hospitalized patients, for example) increases the potential to generalize the results to broader populations. Results from the Medicare data will be generalizable to US patients with T2DM aged 65 years or older and not in a residential care facility. Results from HIRD^{SM} data will be generalizable to the patients with T2DM among the employable US population.

6 STUDY CONDUCT

This study will be conducted in accordance with International Society for Pharmacoepidemiology (ISPE) (2015) *Guidelines for Good Epidemiology Practices* and applicable regulatory requirements including European Medicines Agency (EMA) *Guideline on Good Pharmacovigilance Practices: Module VIII – Post-Authorisation Safety Studies* (EMA, 2016). As with all research at RTI International, RTI-HS will request review of the protocol by the RTI International institutional review board (IRB), and we anticipate that the IRB will exempt the protocol from IRB review because the study data will not have any patient identifiers.

6.1 Ethics Committee Review and Informed Consent

This study does not require review and approval by ethics committees or informed consent.

6.1.1 Ethics Committee Review

6.1.1.1 CPRD

RTI-HS will prepare the request and submit the study protocols to the CPRD's Independent Scientific Advisory Committee (ISAC) (http://www.CPRD.com/isac) for approval. The CPRD has obtained ethical approval from a Multicenter Research Ethics Committee (MREC) for all observational research using CPRD data without patient involvement; however, ISAC may recommend that the MREC review the study documentation if any ethical issues arise.

6.1.1.2 PHARMO

Research using the PHARMO Database Network is conducted in collaboration with investigators at the PHARMO Institute, which conducts research according to the latest directives regarding privacy and handling of data. Ethical approval will not be relevant because pharmacy records and all data sources used are anonymous and are linked through probabilistic linkage using demographic variables of the patients. All other identifying information will be deleted after the linkage with the hospital records from the various data sources. This approach is approved by the Dutch Data Protection Authority. Researchers have information only on sex and age of the patient. Permission is needed to obtain the data from the partnership data sources (NCR and PALGA). Approval for a study using the PHARMO GP data is required from the "Raad van Toezicht," a research ethics review board.

*6.1.1.3 HIRD^SM^*

This component of the larger study is designed as an analysis based on medical and pharmacy claims data from a large insured population in the US. There is no active enrollment or active follow-up of study patients, and no data will be collected directly from individuals.

HealthCore maintains Data Sharing Agreements and Business Associate Agreements with all covered entities who provide data to the HIRD^{SM}. HealthCore's access, use, and disclosure of protected health information (PHI) are in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (45 CFR Part 160 and Subparts A and E of Part 164). HealthCore does not access, use, or disclose identifiable PHI unless under a specific waiver of authorization (e.g., a HIPAA Waiver of Authorization from an IRB). HealthCore accesses the data in a manner that complies with federal and state laws and regulations, including those related to the privacy and security of individually identifiable health information.

As PHI must be accessed in order to conduct the medical record acquisition, a HIPAA Waiver of Authorization will be applied for from an IRB. HealthCore will submit the protocol to a central IRB for review and approval. Approval is typically provided within 2 to 3 weeks of submission. Once IRB approval is obtained, HealthCore's vendor will proceed with medical record acquisition. If changes to the protocol are required, HealthCore will submit an amendment to the IRB. As the IRB is independent, HealthCore cannot control the approval or whether there are conditions for the approval.

Notwithstanding receipt of approval from a central IRB, in some instances, individual institutions may require approval from their local IRB, which would require a separate protocol submission and, in some cases, additional fees. In these cases, HealthCore, RTI-HS, and AZ will need to agree whether or not to proceed with chart acquisition from these institutions.

HealthCore will provide to the vendor only the minimum amount of patient information that is necessary to accomplish the medical record acquisition. HealthCore uses only vendors that follow federal and state laws and regulations, including but not limited to privacy and security rules such as HIPAA.

At no time during the conduct of this study will HealthCore provide patient- or provider-identifying information to RTI-HS, BMS, or AZ. Only aggregated data will be reported to RTI-HS, BMS, and AZ.

6.1.1.4 Medicare

For use of Medicare data, CMS requires that IRB review and approval be obtained before use of Medicare data for research can be approved. This protocol will be reviewed by the RTI International IRB before applying to use Medicare data and will undergo a continuing IRB review at least once per year.

Under the Privacy Rule (CFR 45 164.512), CMS may disclose protected health information for research without documentation of individual authorization only if an IRB or a CMS Privacy Board has approved a waiver of research. Such a waiver must be provided to CMS.

Data requests for research identifiable data must be reviewed by the CMS Privacy Board to ensure that any study patient's privacy is protected and the need for identifiable data is justified.

Investigators will ensure the confidentiality of individually identifiable medical information of the study patients. All personal identifiers will be removed in accordance with applicable laws and regulations from all verification records and files that are accessible to nonstudy personnel, and code keys will be stored separate from the study verification files. All personnel with access to data containing personal identifiers will sign a pledge to maintain the confidentiality of study patients and will maintain an ability to verify the origin and integrity of data sets from which personal identifiers will have been removed.

6.2 Responsibilities Within the Study

The study shall be conducted as described in this approved protocol. All revisions to the protocol must be discussed with, and be prepared by AZ.

6.2.1 Sponsor Roles and Responsibilities

The sponsor, AZ, is responsible for providing reasonable resources for study implementation and to assure study progress. They are also responsible for communicating with regulatory agencies about the study protocol, the progress of the study, and study findings.

6.2.2 Investigator Roles and Responsibilities

The study investigators at RTI-HS, PHARMO, and HealthCore share responsibility with BMS and AZ for the design of the study. The investigators at RTI-HS are responsible for conducting the CPRD and Medicare components in a manner that meets regulatory and methodologic standards, conducting analyses, and preparing scientific reports. The investigators at HealthCore and PHARMO are responsible for analysis in their respective databases in a manner that meets methodologic and regulatory standards, conducting analyses, and preparing scientific reports.

The study shall be conducted as described in the approved protocol. The authors will not develop or implement any deviation or change to the protocol without prior review by AZ.

6.3 Confidentiality of Study Data

The confidentiality of records that could identify patients within the data source must be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).

Data that could directly identify the patient will not be collected in the “study database.”

6.4 Quality Control

Experienced RTI-HS programmers in the United States will perform all analyses for the CPRD data. To ensure the integrity and quality of the study results, RTI-HS will follow the programming validation life cycle process for all analyses. This includes quality-checking programs, logs, and output for accuracy according to relevant standard operating procedures. All programs will be independently reviewed by a second programmer/analyst.

At PHARMO, all researches and analyses are administered in such a manner that the data selection and statistical analyses can be reproduced and verified. All programming will be independently reviewed by an experienced analyst at PHARMO, and all results and reports are audited by the quality-control department. Requests for control of the working methods by external parties need to be sufficiently grounded but can be submitted to the board of directors.

HealthCore’s quality system is organized around the Quality Manual, the quality checks within the project life cycle, and the standard operating procedures. HealthCore performs internal audits to endure adherence to the quality system according to a formal procedure and has procedures for retention of PHI and project data. The study will be tracked at various levels to help ensure that all aspects including project delivery, infrastructure, quality processes, resource management, and financial issues are addressed. To help ensure the highest level of quality on every project, HealthCore has established multiple layers of quality assurance throughout the project life cycle.

- Role-Based Control Checks: Each member of the team is responsible for performing thorough quality assurance checks on his or her work. In addition, the Project Director in collaboration with the Lead Epidemiologist is also accountable for the quality of all deliverables.
- Quality Check Points: Centralized “check points” have been implemented during the data management cycle to help ensure accurate translation of programming requests.
- Quality Assurance Standards: Standard review procedures have been developed and are applied throughout the project life cycle.
- Automation: HealthCore has developed standard definitions of many variables and disease states and developed programs to apply these standards as needed on projects. These standards help ensure consistency, repeatability, and accuracy for each project.

HealthCore’s research team documents the progress and scientific and quality review of all study activities and deliverables (e.g., protocol, reports, and manuscripts) in a project log. The project log provides documentation of the major study tasks related to a specific study activity performed by the research team, to develop and execute the requirements of the protocol or other guiding document for a HealthCore research project. In addition, the project log documents the quality assurance measures performed for each study activity during the conduct of the research project. Also, any research team and/or sponsor interaction resulting in a change to study specifications (e.g., protocol, study database, variables in the analytic files) is described in the project log. This is necessary to ensure that such communications are appropriately documented, that the most up-to-date versions of relevant documents are readily identifiable, and that affected documents are clearly tracked in the project log.

This project will be guided by a written analysis plan to ensure that all collaborators conduct quality-control checks of all aspects of data manipulation and analysis and preparation of study deliverables. The analysis plan will specify that all collaborators will establish and maintain adequate documentation of performance of major tasks. The RTI-HS Office of Quality Assurance will conduct periodic audits during the study to ensure that such documentation meets the necessary standards, especially the completion of these quality-control checks, according to the analysis plan.

6.5 Database Retention and Archiving of Study Documents

The investigator must retain all study records and source documents for the maximum period required by applicable regulations and guidelines, or institution procedures, or for the period specified by the sponsor, whichever is longer. The investigator must contact the sponsor prior to destroying any records associated with the study. Location of the study database and supporting documentation will be outlined in the final observational study report.

The location of analysis data sets and supporting documentation will be outlined in the final observational study report.

6.6 Registration of Study on Public Website

The study was registered in the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) EU PAS Register (ENCePP, 2016) and ClinicalTrials.gov, before the study implementation commenced. The research team and study sponsor adhere to the general principles of transparency and independence in the ENCePP code of conduct (ENCePP, 2014).

6.7 Plans for Disseminating and Communicating Study Results

In accordance with the *Guidelines for Good Pharmacoepidemiology Practices* (ISPE, 2015), there is an ethical obligation to disseminate findings of potential scientific or public health importance, e.g., results pertaining to the safety of a marketed medication. The Consolidated Standards of Reporting Trials (CONSORT) statement refers to randomized studies, but also provides useful guidance applicable to reporting results of nonrandomized studies (Moher et al., 2001). A well-developed publication strategy is encouraged in the Guideline on Good Pharmacovigilance Practices, module VIII, Section B.7 (EMA, 2016).

Reports will be provided after each of the analyses, i.e., the descriptive analysis and the comparative analyses. Personnel at RTI-HS, PHARMO, and HealthCore reserve the right to submit the results from these analyses for publication, as agreed together, and commit that they will publish at least the final results. The authorship of publications shall be in accordance with standards as described in the *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals* (International Committee of Medical Journal Editors, 2016).

7 ADVERSE EVENT REPORTING

7.1 Adverse Event Definitions

An adverse event (AE) is any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment. An AE can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding, for example), symptom, or disease temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product.

A non-serious adverse event is any AE that is not classified as serious.

A serious AE (SAE) is any untoward medical occurrence that at any dose:

- Results in death
- Is life-threatening (defined as an event in which the patient was at risk of death at the time of the event; it does not refer to an event which hypothetically might have caused death if it were more severe)
- Requires inpatient hospitalization or causes prolongation of existing hospitalization (See Note below)
- Results in persistent or significant disability/incapacity

- Is a congenital anomaly/birth defect
- Is an important medical event (defined as a medical event(s) that may not be immediately life-threatening or result in death or hospitalization but, based upon appropriate medical and scientific judgment, may jeopardize the patient or may require intervention [e.g., medical, surgical] to prevent one of the other serious outcomes listed in the definition above.) Examples of such events include, but are not limited to, intensive treatment in an emergency department or at home for allergic bronchospasm; blood dyscrasias or convulsions that do not result in hospitalization.)

Suspected transmission of an infectious agent, pathogenic or nonpathogenic, via the AZ product under study is an SAE.

An ***overdose*** is defined as the accidental or intentional administration of any dose of a product that is considered both excessive and medically important.

Although overdose and cancer are not always serious by regulatory definition, these events are handled as SAEs.

The following hospitalizations are not considered SAEs in AZ studies:

- A visit to the emergency department or other hospital department < 24 hours, that does not result in admission (unless considered an important medical or life-threatening event).
- Elective surgery, planned prior to signing consent.
- Routine health assessment requiring admission for baseline/trending of health status (e.g., routine colonoscopy).
- Medical/surgical admission other than to remedy ill health and planned prior to entry into the study.
- Admission encountered for another life circumstance that carries no bearing on health status and requires no medical/surgical intervention (e.g., lack of housing, economic inadequacy, caregiver respite, family circumstances, administrative reasons).

7.2 Adverse Event Collection and Reporting

All AEs collected will be reported in aggregate in the final study report.

8 GLOSSARY OF TERMS AND LIST OF ABBREVIATIONS

8.1 Glossary of Terms

Not applicable.

8.2 List of Abbreviations

Term	Definition
AD	antidiabetic drug
AE	adverse event
ATC	Anatomical Therapeutic Chemical (classification system)
AZ	AstraZeneca

Term	Definition
BMI	body mass index
BMS	Bristol-Myers Squibb
BRCA1	abbreviation for a human gene: breast cancer 1, early onset
BRCA2	abbreviation for a human gene: breast cancer 2, early onset
CI	confidence interval
CL	confidence limit
CMS	Centers for Medicare and Medicaid Services
CONSORT	Consolidated Standards of Reporting Trials
CPRD	Clinical Practice Research Datalink
CPT	Current Procedural Terminology (coding system)
CVOT	cardiovascular outcome trial
DAPA	Dapagliflozin
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
GP	general practitioner
GPRD	General Practice Research Database, now the CPRD
HCPCS	Healthcare Common Procedure Coding System
HES	Hospital Episode Statistics
HIPAA	Health Insurance Portability and Accountability Act
HIRD^{SM}	HealthCore Integrated Research Database
HIV	human immunodeficiency virus
HR	hazard ratio
ICD-9	International Classification of Diseases, 9th Edition
ICD-9-CM	International Classification of Diseases, 9th Edition, Clinical Modification
ICD-O-3	International Classification of Diseases for Oncology, Third Edition
IRB	institutional review board
IRR	incidence rate ratio
ISAC	Independent Scientific Advisory Committee
ISPE	International Society for Pharmacoepidemiology
MREC	Multicenter Research Ethics Committee
NCR	Netherlands Cancer Registry
NDC	National Drug Code
NEC	not elsewhere classified
NHL	non-Hodgkin lymphoma

Term	Definition
NOS	not otherwise specified
OS	otherwise specified
PALGA	Dutch National Pathology Registry
PASS	postauthorization safety study
PHARMO	PHARMO Database Network, the Netherlands
PHI	protected health information
RR	relative risk
RTI-HS	RTI Health Solutions
SAE	serious adverse event
SAP	statistical analysis plan
SGLT2	sodium-glucose cotransporter 2
T1DM	type 1 diabetes mellitus
T2DM	type 2 diabetes mellitus
UK	United Kingdom
US	United States
WHO	World Health Organization

9 REFERENCES

American Diabetes Association. Standards of medical care in diabetes-2014. Diabetes Care. 2014 Jan;37 Suppl 1:S14-80.

Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107.

Boccuzzi SJ, Wogen J, Fox J, Sung JCY, Shah AB, Kim J. Utilization of oral hypoglycemic agents in a drug-insured U.S. population. Diabetes Care. 2001;24:1411-15.

Bodmer M, Meier C, Krähenbühl S, Jick SS, Meier CR. Long-term metformin use is associated with decreased risk of breast cancer. Diabetes Care. 2010;33:1304-8.

Braitman LE, Rosenbaum PR. Rare outcomes, common treatments : Analytic strategies using propensity scores. Ann Intern Med. 2002;137:693-5.

Bristol-Meyers Squibb (BMS) and AstraZeneca (AZ). Background document: dapagliflozin. BMS-512148. NDA 202293. 13 June 2011. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/EndocrinologicandMetabolicDrugsAdvisoryCommittee/UCM262996.pdf. Accessed 22 February 2017.

Bristol-Myers Squibb and AstraZeneca. Background document: dapagliflozin, BMS-512148, NDA 202293. Advisory Committee briefing materials. 04 November 2013. Available at:

http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/EndocrinologicandMetabolicDrugsAdvisoryCommittee/UCM378079.pdf. Accessed 22 February 2017.

Centers for Medicare and Medicaid Services (CMS). Medicare Program - General Information. 2013. Available at: http://www.cms.gov/Medicare/Medicare-General-Information/MedicareGenInfo/index.html. Accessed 22 February 2017.

Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158:280-7.

Chang HY, Weiner JP, Richards TM, Bleich SN, Segal JB. Validating the adapted Diabetes Complications Severity Index in claims data. Am J Manag Care. 2012 Nov;18(11):721-6.

Chlebowski RT, McTiernan A, Wactawski-Wende J, Manson JE, Aragaki AK, Rohan T, et al. Diabetes, metformin, and breast cancer in postmenopausal women. J Clin Oncol. 2012;30:2844-52.

Currie CJ, Poole CD, Gale EAM. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. Diabetologia. 2009;52:1766-77.

D’Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17:2265-81.

Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013 Apr;49(6):1374-403.

European Medicines Agency. Guideline on good pharmacovigilance practices (GVP). Module VIII – Post-authorisation safety studies. (Rev 2). European Medicines Agency; 04 August 2016. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf. Accessed 24February 2017.

European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). The ENCePP code of conduct for scientific independence and transparency in the conduct of pharmacoepidemiological and pharmacovigilance studies. Revision 3. 21 February 2014. Available at: http://www.encepp.eu/code_of_conduct/documents/ENCePPCodeofConduct_Rev3.pdf. Accessed 22 February 2017.

European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). The European Union Electronic Register of Post-Authorisation Studies (EU PAS Register). 15 July 2016. Available at: http://www.encepp.eu/encepp_studies/indexRegister.shtml. Accessed 22 February 2017.

Gelfand JM, Margolis DJ, Dattani H. The UK General Practice Research Database. In: Strom BL, editor Pharmacoepidemiology, 4th edition. John Wiley & Sons; 2005. p. 337-46.

Giovannucci E, Harlan DM, Archer MC, Bergenstal RM, Gapstur ^{SM}, Habel LA, et al. Diabetes and cancer: a consensus report. CA Cancer J Clin. 2010;60:207-21.

Hall GC, McMahon AD, Carroll D, Home PD. Macrovascular and microvascular outcomes after beginning of insulin versus additional oral glucose-lowering therapy in people with type 2 diabetes: an observational study. Pharmacoepidemiol Drug Saf. 2012;21:305-13.

Herings RMC, Pedersen L. Pharmacy-based medical record linkage systems. In: Strom BL, Kimmel SE, Hennessy S, editors. Pharmacoepidemiology, 5th Edition. John Wiley & Sons, Ltd; 2012. p. 270-86.

International Committee of Medical Journal Editors (ICMJE). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. December 2016. Available at: http://www.icmje.org/urm_main.html. Accessed 24 February 2017.

International Society for Pharmacoepidemiology (ISPE). Guidelines for good pharmacoepidemiology practices (GPP). Revision 3. June 2015. Available at: http://www.pharmacoepi.org/resources/guidelines_08027.cfm. Accessed 24 February 2017.

Jick H, Jick S, Derby LE, Vasilakis C, Myers MW, Meier CR. Calcium-channel blockers and risk of cancer. Lancet. 1997;349:525-8.

Kasper JS, Giovannucci E. A meta-analysis of diabetes mellitus and the risk of prostate cancer. Cancer Epidemiol Biomarkers Prev. 2006;15:2056-62.

Kasper JS, Liu Y, Giovannucci E. Diabetes mellitus and risk of prostate cancer in the health professionals follow-up study. Int J Cancer. 2009;124:1398-403.

Kaye JA, Derby LE, Melero-Montes MM, Quinn M, Jick H. The incidence of breast cancer in the General Practice Research Database compared with national cancer registration data. Br J Cancer. 2000;83:1556-8.

Kaye JA, Myers MW, Jick H. Acetaminophen and the risk of renal and bladder cancer in the General Practice Research Database. Epidemiology. 2001;12:690-4.

Kirkali Z, Chan T, Manoharan M, Algaba F, Busch C, Cheng L, et al. Bladder cancer: epidemiology, staging and grading, and diagnosis. Urology. 2005;66(suppl 6A):4-34.

Larsson SC, Orsini N, Brismar K, Wolk A. Diabetes mellitus and risk of bladder cancer: a meta-analysis. Diabetologia. 2006;49:2819-23.

Larsson SC, Mantzoros CS, Wolk A. Diabetes mellitus and risk of breast cancer: a meta-analysis. Int J Cancer. 2007;121:856-62.

Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology. 2003;14:451-8.

Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Dordrecht Heidelberg London New York: Springer; 2009.

Lewis JD, Ferrara A, Peng T, Hedderson M, Bilker WB, Quesenberry CP, et al. Risk of bladder cancer among diabetic patients treated with pioglitazone. Diabetes Care. 2011;34:916-22.

Malone KE. Diethylstilbestrol (DES) and breast cancer. Epidemiol Rev 1993;15:108-9.

Matanoski GM, Elliott EA. Bladder cancer epidemiology. Epidemiol Rev. 1981;3:203-29.

Michels KB, Solomon CG, Hu FB, Rosner BA, Hankinson SE, Colditz GA, et al. Type 2 diabetes and subsequent incidence of breast cancer in the Nurses' Health Study. Diabetes Care 2003;26:1752-8.

Moher D, Schulz KF, Altman D; CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285:1987-91.

National Institute for Health and Clinical Excellence. Type 2 diabetes in adults: management. 2015. Available at: https://www.nice.org.uk/guidance/ng28. Accessed 22 February 2017.

Neumann A, Weill A, Ricordeau P, Fagot JP, Alla F, Allemand H. Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study. Diabetologia. 2012;55:1953-62.

Perkins ^{SM}, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiol Drug Saf. 2000;9:93-101.

Rothman KJ. Episheet: spreadsheets for the analysis of epidemiologic data [software]. 04 October 2012. Available at: http://krothman.hostbyet2.com/Episheet.xls. Accessed 24 February 2017.

Ruiter R, Visser LE, van Herk-Sukel MP, Coebergh JW, Haak HR, Geelhoed-Duijvestijn PH, et al. Risk of cancer in patients on insulin glargine and other insulin analogues in comparison with those on human insulin: results from a large population-based follow-up study. Diabetologia. 2012;55:51-62.

van Herk-Sukel MP, van de Poll-Franse LV, Lemmens VE, Vreugdenhil G, Pruijt JF, Coebergh JW, et al. New opportunities for drug outcomes research in cancer patients: the linkage of the Eindhoven Cancer Registry and the PHARMO Record Linkage System. Eur J Cancer. 2010;46:395-404.

Vigneri P, Frasca, F, Sciacca L, et al. Diabetes and cancer. Endocr Relat Cancer. 2009;16:1103-23.

Walker AM. Identification of esophageal cancer in the General Practice Research Database. Pharmacoepi Drug Saf. 2011;20:1159-67.

Wei L, MacDonald TM, Mackenzie IS. Pioglitazone and bladder cancer: a propensity score matched cohort study. Br J Clin Pharmacol. 2013 Jan;75(1):254-9.

Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. JAMA. 2007;298:2654-64.

APPENDIX 1. CLINICAL READ CODES FOR CANCERS TO BE STUDIED

Table 1-1: Clinical Read Codes for Female Breast Cancer

Read Codes	Description
B34…00	Malignant neoplasm of female breast
B34…11	Cancer of female breast
B340.00	Malignant neoplasm of nipple and areola of female breast
B340000	Malignant neoplasm of nipple of female breast
B340100	Malignant neoplasm of areola of female breast
B340z00	Malignant neoplasm of nipple or areola of female breast NOS
B341.00	Malignant neoplasm of central part of female breast
B342.00	Malignant neoplasm of upper-inner quadrant of female breast
B343.00	Malignant neoplasm of lower-inner quadrant of female breast
B344.00	Malignant neoplasm of upper-outer quadrant of female breast
B345.00	Malignant neoplasm of lower-outer quadrant of female breast
B346.00	Malignant neoplasm of axillary tail of female breast
B347.00	Malignant neoplasm, overlapping lesion of breast
B34y.00	Malignant neoplasm of other site of female breast
B34y000	Malignant neoplasm of ectopic site of female breast
B34yz00	Malignant neoplasm of other site of female breast NOS
B34z.00	Malignant neoplasm of female breast NOS
Byu6.00	[X]Malignant neoplasm of breast
BB94.00	[M]Juvenile breast carcinoma
BB94.11	[M]Secretory breast carcinoma
BB9J.11	[M]Paget's disease, breast
BB9K.00	[M]Paget's disease and infiltrating breast duct carcinoma
BB9K000	[M]Paget's disease and intraductal carcinoma of breast
BB5K.00	[M]Cribriform carcinoma
BB91000	[M]Intraductal papillary adenocarcinoma with invasion
BB91.00	[M]Infiltrating duct carcinoma
BB9G.00	[M]Infiltrating ductular carcinoma
BB91100	[M]Infiltrating duct and lobular carcinoma
BB93.00	[M]Comedocarcinoma NOS
BB9F.00	[M]Lobular carcinoma NOS
BB9H.00	[M]Inflammatory carcinoma
BB9J.00	[M]Paget's disease, mammary

Read Codes	Description
BBM9.00	[M]Cystosarcoma phyllodes, malignant

NOS = not otherwise specified.

Table 1-2: Clinical Read Codes for Bladder Cancer

Read Codes	Description
B49…00	Malignant neoplasm of urinary bladder
B490.00	Malignant neoplasm of trigone of urinary bladder
B491.00	Malignant neoplasm of dome of urinary bladder
B492.00	Malignant neoplasm of lateral wall of urinary bladder
B493.00	Malignant neoplasm of anterior wall of urinary bladder
B494.00	Malignant neoplasm of posterior wall of urinary bladder
B495.00	Malignant neoplasm of bladder neck
B496.00	Malignant neoplasm of ureteric orifice
B497.00	Malignant neoplasm of urachus
B49y.00	Malignant neoplasm of other site of urinary bladder
B49y000	Malignant neoplasm, overlapping lesion of bladder
B49z.00	Malignant neoplasm of urinary bladder NOS
B49z.00	Malignant neoplasm of urinary bladder NOS
B837.00	Carcinoma in situ of bladder
BB4..00	[M]Transitional cell papillomas and carcinomas
BB43.00	[M]Transitional cell carcinoma NOS
BB43.11	[M]Urothelial carcinoma
BB47.00	[M]Transitional cell carcinoma, spindle cell type
BB4A.00	[M]Papillary transitional cell carcinoma
BB4B.00	[M]Grade 1 (Stage pTa) papillary urothelial/transit cell ca
BB4C.00	[M]Grade 2 (Stage pTa) papillary urothelial/transit cell ca
BB4D.00	[M]Grade 3 (Stage pTa) papillary urothelial/transit cell ca

NOS = not otherwise specified.

Table 1-3: Clinical Read Codes for Colon/Rectum Cancer

Read Code	Description
B13…00	Malignant neoplasm of colon
B130.00	Malignant neoplasm of hepatic flexure of colon
B131.00	Malignant neoplasm of transverse colon
B132.00	Malignant neoplasm of descending colon

Read Code	Description
B133.00	Malignant neoplasm of sigmoid colon
B134.00	Malignant neoplasm of caecum
B134.11	Carcinoma of caecum
B135.00	Malignant neoplasm of appendix
B136.00	Malignant neoplasm of ascending colon
B137.00	Malignant neoplasm of splenic flexure of colon
B138.00	Malignant neoplasm, overlapping lesion of colon
B139.00	Hereditary nonpolyposis colon cancer
B13y.00	Malignant neoplasm of other specified sites of colon
B13z.00	Malignant neoplasm of colon NOS
B13z.11	Colonic cancer
B14…00	Malignant neoplasm of rectum, rectosigmoid junction and anus
B140.00	Malignant neoplasm of rectosigmoid junction
B141.00	Malignant neoplasm of rectum
B141.11	Carcinoma of rectum
B141.12	Rectal carcinoma
B14y.00	Malig neop other site rectum, rectosigmoid junction and anus
B14z.00	Malignant neoplasm rectum, rectosigmoid junction and anus NOS
B18y200	Malignant neoplasm of mesorectum
BB5L100	[M]Adenocarcinoma in adenomatous polyp
BB5L300	[M]Adenocarcinoma in multiple adenomatous polyps
BB5N100	[M]Adenocarcinoma in adenomatous polyposis coli
BB5R100	[M]Carcinoid tumor, malignant
BB5R500	[M]Carcinoid tumor, nonargentaffin, malignant
BB5U100	[M]Adenocarcinoma in villous adenoma
BB5U200	[M]Villous adenocarcinoma

NOS = not otherwise specified.

Table 1-4: Clinical Read Codes for Corpus Uteri Cancer

Read Code	Description
B40..00	Malignant neoplasm of uterus, part unspecified
B4300	Malignant neoplasm of body of uterus
B430.00	Malignant neoplasm of corpus uteri, excluding isthmus
B430000	Malignant neoplasm of cornu of corpus uteri
B430100	Malignant neoplasm of fundus of corpus uteri

Read Code	Description
B430200	Malignant neoplasm of endometrium of corpus uteri
B430211	Malignant neoplasm of endometrium
B430300	Malignant neoplasm of myometrium of corpus uteri
B430z00	Malignant neoplasm of corpus uteri NOS
B431.00	Malignant neoplasm of isthmus of uterine body
B431000	Malignant neoplasm of lower uterine segment
B431z00	Malignant neoplasm of isthmus of uterine body NOS
B432.00	Malignant neoplasm of overlapping lesion of corpus uteri
B43y.00	Malignant neoplasm of other site of uterine body
B43z.00	Malignant neoplasm of body of uterus NOS

NOS = not otherwise specified.

Table 1-5: Clinical Read Codes for Lung Cancer

Read Code	Description
B22…00	Malignant neoplasm of trachea, bronchus and lung
B220.00	Malignant neoplasm of trachea
B220100	Malignant neoplasm of mucosa of trachea
B220z00	Malignant neoplasm of trachea NOS
B221.00	Malignant neoplasm of main bronchus
B221000	Malignant neoplasm of carina of bronchus
B221100	Malignant neoplasm of hilus of lung
B221z00	Malignant neoplasm of main bronchus NOS
B222.00	Malignant neoplasm of upper lobe, bronchus or lung
B222.11	Pancoast's syndrome
B222000	Malignant neoplasm of upper lobe bronchus
B222100	Malignant neoplasm of upper lobe of lung
B222z00	Malignant neoplasm of upper lobe, bronchus or lung NOS
B223.00	Malignant neoplasm of middle lobe, bronchus or lung
B223000	Malignant neoplasm of middle lobe bronchus
B223100	Malignant neoplasm of middle lobe of lung
B223z00	Malignant neoplasm of middle lobe, bronchus or lung NOS
B224.00	Malignant neoplasm of lower lobe, bronchus or lung
B224000	Malignant neoplasm of lower lobe bronchus
B224100	Malignant neoplasm of lower lobe of lung
B224z00	Malignant neoplasm of lower lobe, bronchus or lung NOS

Read Code	Description
B225.00	Malignant neoplasm of overlapping lesion of bronchus & lung
B22y.00	Malignant neoplasm of other sites of bronchus or lung
B22z.00	Malignant neoplasm of bronchus or lung NOS
B22z.11	Lung cancer
BB1J.00	[M]Small cell carcinoma NOS
BB1J.12	[M]Round cell carcinoma
BB1K.00	[M]Oat cell carcinoma
BB1L.00	[M]Small cell carcinoma, fusiform cell type
BB1M.00	[M]Small cell carcinoma, intermediate cell
BB1N.00	[M]Small cell-large cell carcinoma
BB1P.00	[M]Non-small cell carcinoma
BB5R111	[M]Carcinoid bronchial adenoma
BB5S200	[M]Bronchiolo-alveolar adenocarcinoma
BB5S211	[M]Alveolar cell carcinoma
BB5S212	[M]Bronchiolar carcinoma
BB5S400	[M]Alveolar adenocarcinoma
BBLM.00	[M]Pulmonary blastoma
BBTL.00	[M]Intravascular bronchial alveolar tumor
Byu2000	[X]Malignant neoplasm of bronchus or lung, unspecified
Byu5011	[X]Mesothelioma of lung

NOS = not otherwise specified.

Table 1-6: Clinical Read Codes for Melanoma of Skin

Read Code	Description
4M3..00	Breslow depth staging for melanoma
4M71.00	Clark melanoma level 2
4M72.00	Clark melanoma level 3
4M73.00	Clark melanoma level 4
4M74.00	Clark melanoma level 5
B32…00	Malignant melanoma of skin
B320.00	Malignant melanoma of lip
B321.00	Malignant melanoma of eyelid including canthus
B322.00	Malignant melanoma of ear and external auricular canal
B322000	Malignant melanoma of auricle (ear)
B322100	Malignant melanoma of external auditory meatus

Read Code	Description
B322z00	Malignant melanoma of ear and external auricular canal NOS
B323.00	Malignant melanoma of other and unspecified parts of face
B323000	Malignant melanoma of external surface of cheek
B323100	Malignant melanoma of chin
B323200	Malignant melanoma of eyebrow
B323300	Malignant melanoma of forehead
B323400	Malignant melanoma of external surface of nose
B323500	Malignant melanoma of temple
B323z00	Malignant melanoma of face NOS
B324.00	Malignant melanoma of scalp and neck
B324000	Malignant melanoma of scalp
B324100	Malignant melanoma of neck
B324z00	Malignant melanoma of scalp and neck NOS
B325.00	Malignant melanoma of trunk (excluding scrotum)
B325000	Malignant melanoma of axilla
B325100	Malignant melanoma of breast
B325200	Malignant melanoma of buttock
B325300	Malignant melanoma of groin
B325400	Malignant melanoma of perianal skin
B325500	Malignant melanoma of perineum
B325600	Malignant melanoma of umbilicus
B325700	Malignant melanoma of back
B325800	Malignant melanoma of chest wall
B325z00	Malignant melanoma of trunk, excluding scrotum, NOS
B326.00	Malignant melanoma of upper limb and shoulder
B326000	Malignant melanoma of shoulder
B326100	Malignant melanoma of upper arm
B326200	Malignant melanoma of fore-arm
B326300	Malignant melanoma of hand
B326400	Malignant melanoma of finger
B326500	Malignant melanoma of thumb
B326z00	Malignant melanoma of upper limb or shoulder NOS
B327.00	Malignant melanoma of lower limb and hip
B327000	Malignant melanoma of hip

Read Code	Description
B327100	Malignant melanoma of thigh
B327200	Malignant melanoma of knee
B327300	Malignant melanoma of popliteal fossa area
B327400	Malignant melanoma of lower leg
B327500	Malignant melanoma of ankle
B327600	Malignant melanoma of heel
B327700	Malignant melanoma of foot
B327800	Malignant melanoma of toe
B327900	Malignant melanoma of great toe
B327z00	Malignant melanoma of lower limb or hip NOS
B32y.00	Malignant melanoma of other specified skin site
B32y000	Overlapping malignant melanoma of skin
B32z.00	Malignant melanoma of skin NOS
BBE..00	[M]Naevi and melanomas
BBE1.00	[M]Malignant melanoma NOS
BBE1.11	[M]Melanocarcinoma
BBE1.12	[M]Melanoma NOS
BBE1.14	[M]Naevocarcinoma
BBE1000	[M]Malignant melanoma, regressing
BBE1100	[M]Desmoplastic melanoma, malignant
BBE2.00	[M]Nodular melanoma
BBE4.00	[M]Balloon cell melanoma
BBEA.00	[M]Amelanotic melanoma
BBEC.00	[M]Malignant melanoma in junctional naevus
BBEG.00	[M]Malignant melanoma in Hutchinson's melanotic freckle
BBEG.11	[M]Lentigo maligna melanoma
BBEG000	[M]Acral lentiginous melanoma, malignant
BBEH.00	[M]Superficial spreading melanoma
BBEM.00	[M]Malignant melanoma in giant pigmented naevus
BBEP.00	[M]Epithelioid cell melanoma
BBEQ.00	[M]Spindle cell melanoma NOS
BBES.00	[M]Spindle cell melanoma, type B
BBET.00	[M]Mixed epithelioid and spindle melanoma
BBEV.00	[M]Blue naevus, malignant

Read Code	Description
BBEz.00	[M]Naevi or melanoma NOS
Byu4.00	[X]Melanoma and other malignant neoplasms of skin
Byu4000	[X]Malignant melanoma of other+unspecified parts of face
Byu4100	[X]Malignant melanoma of skin, unspecified

NOS = not otherwise specified.

Table 1-7: Clinical Read Codes for Non-Hodgkin Lymphoma

Read Code	Description
B60..00	Lymphosarcoma and reticulosarcoma
B600.00	Reticulosarcoma
B600000	Reticulosarcoma of unspecified site
B600100	Reticulosarcoma of lymph nodes of head, face and neck
B600300	Reticulosarcoma of intra-abdominal lymph nodes
B600700	Reticulosarcoma of spleen
B600z00	Reticulosarcoma NOS
B601.00	Lymphosarcoma
B601000	Lymphosarcoma of unspecified site
B601100	Lymphosarcoma of lymph nodes of head, face and neck
B601200	Lymphosarcoma of intrathoracic lymph nodes
B601300	Lymphosarcoma of intra-abdominal lymph nodes
B601500	Lymphosarcoma of lymph nodes of inguinal region and leg
B601700	Lymphosarcoma of spleen
B601800	Lymphosarcoma of lymph nodes of multiple sites
B601z00	Lymphosarcoma NOS
B602.00	Burkitt's lymphoma
B602100	Burkitt's lymphoma of lymph nodes of head, face and neck
B602200	Burkitt's lymphoma of intrathoracic lymph nodes
B602300	Burkitt's lymphoma of intra-abdominal lymph nodes
B602500	Burkitt's lymphoma of lymph nodes of inguinal region and leg
B602z00	Burkitt's lymphoma NOS
B60y.00	Other specified reticulosarcoma or lymphosarcoma
B60z.00	Reticulosarcoma or lymphosarcoma NOS
B6200	Other malignant neoplasm of lymphoid and histiocytic tissue
B620.00	Nodular lymphoma (Brill - Symmers disease)
B620000	Nodular lymphoma of unspecified site

Read Code	Description
B620100	Nodular lymphoma of lymph nodes of head, face and neck
B620200	Nodular lymphoma of intrathoracic lymph nodes
B620300	Nodular lymphoma of intra-abdominal lymph nodes
B620500	Nodular lymphoma of lymph nodes of inguinal region and leg
B620800	Nodular lymphoma of lymph nodes of multiple sites
B620z00	Nodular lymphoma NOS
B621.00	Mycosis fungoides
B621000	Mycosis fungoides of unspecified site
B621300	Mycosis fungoides of intra-abdominal lymph nodes
B621400	Mycosis fungoides of lymph nodes of axilla and upper limb
B621500	Mycosis fungoides of lymph nodes of inguinal region and leg
B621800	Mycosis fungoides of lymph nodes of multiple sites
B621z00	Mycosis fungoides NOS
B622.00	Sezary's disease
B622z00	Sezary's disease NOS
B62x500	Malignant immunoproliferative small intestinal disease
B627.00	Non-Hodgkin's lymphoma
B627.11	Non-Hodgkin lymphoma
B627000	Follicular non-Hodgkin's small cleaved cell lymphoma
B627100	Follicular non-Hodg mixed sml cleavd & lge cell lymphoma
B627200	Follicular non-Hodgkin's large cell lymphoma
B627300	Diffuse non-Hodgkin's small cell (diffuse) lymphoma
B627400	Diffuse non-Hodgkin's small cleaved cell (diffuse) lymphoma
B627500	Diffuse non-Hodgkin mixed sml & lge cell (diffuse) lymphoma
B627600	Diffuse non-Hodgkin's immunoblastic (diffuse) lymphoma
B627700	Diffuse non-Hodgkin's lymphoblastic (diffuse) lymphoma
B627800	Diffuse non-Hodgkin's lymphoma undifferentiated (diffuse)
B627900	Mucosa-associated lymphoma
B627911	Maltoma
B627A00	Diffuse non-Hodgkin's large cell lymphoma
B627B00	Other types of follicular non-Hodgkin's lymphoma
B627C00	Follicular non-Hodgkin's lymphoma
B627C11	Follicular lymphoma NOS
B627D00	Diffuse non-Hodgkin's centroblastic lymphoma

Read Code	Description
B627E00	Diffuse large B-cell lymphoma
B627F00	Extranod marg zone B-cell lymphom mucosa-assoc lymphoid tiss
B627G00	Mediastinal (thymic) large B-cell lymphoma
B627W00	Unspecified B-cell non-Hodgkin's lymphoma
B627X00	Diffuse non-Hodgkin's lymphoma, unspecified
B628.00	Follicular lymphoma
B628000	Follicular lymphoma grade 1
B628100	Follicular lymphoma grade 2
B628200	Follicular lymphoma grade 3
B628300	Follicular lymphoma grade 3a
B628400	Follicular lymphoma grade 3b
B628500	Diffuse follicle center lymphoma
B628600	Cutaneous follicle center lymphoma
B628700	Other types of follicular lymphoma
B629.00	Multifocal multisystemic dissem Langerhans-cell histiocytosis
B62A.00	Sarcoma of dendritic cells
B62C.00	Unifocal Langerhans-cell histiocytosis
B62D.00	Histiocytic sarcoma
B62E.00	T/NK-cell lymphoma
B62E100	Anaplastic large cell lymphoma, ALK-positive
B62E200	Anaplastic large cell lymphoma, ALK-negative
B62E300	Cutaneous T-cell lymphoma
B62E400	Extranodal NK/T-cell lymphoma, nasal type
B62E500	Hepatosplenic T-cell lymphoma
B62E600	Enteropathy-associated T-cell lymphoma
B62E700	Subcutaneous panniculitic T-cell lymphoma
B62E800	Blastic NK-cell lymphoma
B62E900	Angioimmunoblastic T-cell lymphoma
B62EA00	Primary cutaneous CD30-positive T-cell proliferations
B62Ew00	Other mature T/NK-cell lymphoma
B62F.00	Nonfollicular lymphoma
B62F.11	Nonfollicular lymphoma
B62F000	Small cell B-cell lymphoma
B62F100	Mantle cell lymphoma

Read Code	Description
B62F200	Lymphoblastic (diffuse) lymphoma
B62x.00	Malignant lymphoma otherwise specified
B62x000	T-zone lymphoma
B62x100	Lymphoepithelioid lymphoma
B62x200	Peripheral T-cell lymphoma
B62x400	Malignant reticulosis
B62x500	Malignant immunoproliferative small intestinal disease
B62x600	True histiocytic lymphoma
B62xX00	Oth and unspecif peripheral & cutaneous T-cell lymphomas
B62y.00	Malignant lymphoma NOS
B62y000	Malignant lymphoma NOS of unspecified site
B62y100	Malignant lymphoma NOS of lymph nodes of head, face and neck
B62y200	Malignant lymphoma NOS of intrathoracic lymph nodes
B62y300	Malignant lymphoma NOS of intra-abdominal lymph nodes
B62y400	Malignant lymphoma NOS of lymph nodes of axilla and arm
B62y500	Malignant lymphoma NOS of lymph node inguinal region and leg
B62y600	Malignant lymphoma NOS of intrapelvic lymph nodes
B62y700	Malignant lymphoma NOS of spleen
B62y800	Malignant lymphoma NOS of lymph nodes of multiple sites
B62yz00	Malignant lymphoma NOS
B641.00	Chronic lymphoid leukemia
B641.11	Chronic lymphatic leukemia
B64y200	Adult T-cell leukemia

NOS = not otherwise specified.

Table 1-8: Clinical Read Codes for Ovary Cancer

Read Code	Description
B44…00	Malignant neoplasm of ovary and other uterine adnexa
B440.00	Malignant neoplasm of ovary
B440.11	Cancer of ovary
B441.00	Malignant neoplasm of fallopian tube
B442.00	Malignant neoplasm of broad ligament
B443.00	Malignant neoplasm of parametrium
B44y.00	Malignant neoplasm of other site of uterine adnexa
B44z.00	Malignant neoplasm of uterine adnexa NOS

Read Code	Description
Byu7000	[X]Malignant neoplasm of uterine adnexa, unspecified
BBC4.00	[M]Granulosa cell tumor, malignant
BB81200	[M]Serous cystadenocarcinoma, NOS
BB81500	[M]Papillary cystadenocarcinoma, NOS
BB81800	[M]Papillary serous cystadenocarcinoma
BB81B00	[M]Serous surface papillary carcinoma
BB81E00	[M]Mucinous cystadenocarcinoma NOS
BB81H00	[M]Papillary mucinous cystadenocarcinoma
BBM0100	[M]Brenner tumor, malignant
BBQA100	[M]Struma ovarii, malignant
BB81.11	[M]Ovarian cystadenoma or carcinoma
BBQA200	[M]Strumal carcinoid
BB5R600	[M]Mucocarcinoid tumor, malignant

NOS = not otherwise specified.

Table 1-9: Clinical Read Codes for Prostate Cancer

Read Code	Description
B46…00	Malignant neoplasm of prostate

Table 1-10: Clinical Read Codes for Stomach Cancer

Read Code	Description
B11…00	Malignant neoplasm of stomach
B11..11	Gastric neoplasm
B110.00	Malignant neoplasm of cardia of stomach
B110000	Malignant neoplasm of cardiac orifice of stomach
B110100	Malignant neoplasm of cardio-oesophageal junction of stomach
B110111	Malignant neoplasm of gastro-oesophageal junction
B110z00	Malignant neoplasm of cardia of stomach NOS
B111.00	Malignant neoplasm of pylorus of stomach
B111000	Malignant neoplasm of prepylorus of stomach
B111100	Malignant neoplasm of pyloric canal of stomach
B111z00	Malignant neoplasm of pylorus of stomach NOS
B112.00	Malignant neoplasm of pyloric antrum of stomach
B113.00	Malignant neoplasm of fundus of stomach

Read Code	Description
B114.00	Malignant neoplasm of body of stomach
B115.00	Malignant neoplasm of lesser curve of stomach unspecified
B116.00	Malignant neoplasm of greater curve of stomach unspecified
B117.00	Malignant neoplasm, overlapping lesion of stomach
B118.00	Siewert type II adenocarcinoma
B119.00	Siewert type III adenocarcinoma
B11y.00	Malignant neoplasm of other specified site of stomach
B11y000	Malignant neoplasm of anterior wall of stomach NEC
B11y100	Malignant neoplasm of posterior wall of stomach NEC
B11yz00	Malignant neoplasm of other specified site of stomach NOS
B11z.00	Malignant neoplasm of stomach NOS
BB55.00	[M]Linitis plastica
BB57.00	[M]Adenocarcinoma, intestinal type
BB58.00	[M]Carcinoma, diffuse type
BB5R100	[M]Carcinoid tumor, malignant
BB5R500	[M]Carcinoid tumor, nonargentaffin, malignant

NEC = not elsewhere classified; NOS = not otherwise specified.

APPENDIX 2. ICD-O-3 CODES FOR CANCERS TO BE STUDIED

Table 2-1: ICD-O-3 Topography Codes

Cancer Site	ICD-O-3 Topography Code
Female breast	C50, in a female patient
Bladder	C67
Colon/rectum	C18, C19, or C20
Corpus uteri	C54
Lung	C34
Skin	C44
Ovary	C56
Prostate	C61
Stomach	C16
Trachea	C33

ICD-O-3 = *International Classification of Diseases for Oncology, Third Edition*.

Note: To identify the cancers to be studied, all topography codes must occur in combination with any morphology code (except 9590 to 9989, codes for hematopoietic and lymphoid neoplasms) and with behavior code /2 or /3.

Table 2-2: ICD-O-3 Codes for Melanoma of Skin

Topography code C44 or C80 with morphology code 8720 to 8790 and with behavior code /2 or /3. The resulting morphology/behavior code combinations are as follows:

ICD-O-3 Code	Description
8720/3	Malignant melanoma, NOS
8721/3	Nodular melanoma
8722/3	Balloon cell melanoma
8723/3	Malignant melanoma, regressing
8730/3	Amelanotic melanoma
8740/3	Malignant melanoma in junctional nevus
8741/3	Malignant melanoma in precancerous melanosis
8742/2	Lentigo maligna
8742/3	Lentigo maligna melanoma
8743/3	Superficial spreading melanoma
8744/3	Acral lentiginous melanoma, malignant
8745/3	Desmoplastic melanoma, malignant
8746/3	Mucosal lentiginous melanoma
8761/3	Malignant melanoma in giant pigmented nevus
8770/3	Mixed epithelioid and spindle cell melanoma
8771/3	Epithelioid cell melanoma

Topography code C44 or C80 with morphology code 8720 to 8790 and with behavior code /2 or /3. The resulting morphology/behavior code combinations are as follows:

ICD-O-3 Code	Description
8772/3	Spindle cell melanoma, NOS
8780/3	Blue nevus, malignant

ICD-O-3 = *International Classification of Diseases for Oncology, Third Edition*; NOS = not otherwise specified.

Table 2-3: ICD-O-3 Codes for Non-Hodgkin Lymphoma

No criterion for topography. Morphology/behavior codes are as follows:

ICD-O-3 Code	Description
9591/3	Malignant lymphoma, non-Hodgkin, NOS
9596/3	Composite Hodgkin and non-Hodgkin lymphoma
9670/3	Malignant lymphoma, small B lymphocytic, NOS
9671/3	Malignant lymphoma, lymphoplasmacytic
9673/3	Mantle cell lymphoma
9675/3	Malignant lymphoma, mixed small and large cell, diffuse
9678/3	Primary effusion lymphoma
9679/3	Mediastinal large B-cell lymphoma
9680/3	Malignant lymphoma, large B-cell, diffuse, NOS
9684/3	Malignant lymphoma, large B-cell, diffuse, immunoblastic, NOS
9687/3	Burkitt lymphoma, NOS
9689/3	Splenic marginal zone B-cell lymphoma
9690/3	Follicular lymphoma, NOS
9691/3	Follicular lymphoma, grade 2
9695/3	Follicular lymphoma, grade 1
9698/3	Follicular lymphoma, grade 3
9699/3	Marginal zone B-cell lymphoma, NOS
9700/3	Mycosis fungoides
9701/3	Sezary syndrome
9702/3	Mature T-cell lymphoma, NOS
9705/3	Angioimmunoblastic T-cell lymphoma
9708/3	Subcutaneous panniculitis-like T-cell lymphoma
9709/3	Cutaneous T-cell lymphoma, NOS
9714/3	Anaplastic large cell lymphoma, T-cell and Null cell type
9716/3	Hepatosplenic (gamma-delta) cell lymphoma
9717/3	Intestinal T-cell lymphoma
9718/3	Primary cutaneous CD30+ T-cell lymphoproliferative disorder

No criterion for topography. Morphology/behavior codes are as follows:

ICD-O-3 Code	Description
9719/3	NK/T-cell lymphoma, nasal and nasal type
9727/3	Precursor cell lymphoblastic lymphoma, NOS
9728/3	Precursor B-cell lymphoblastic lymphoma
9729/3	Precursor T-cell lymphoblastic lymphoma
9760/3	Immunoproliferative disease, NOS
9761/3	Waldenstrom macroglobulinemia
9762/3	Heavy chain disease, NOS
9764/3	Immunoproliferative small intestinal disease
9823/3	B-cell chronic lymphocytic leukemia/small lymphocytic lymphoma
9827/3	Adult T-cell leukemia/lymphoma (HTLV-1 positive)

ICD-O-3 = *International Classification of Diseases for Oncology, Third Edition*; NOS = not otherwise specified.

APPENDIX 3. ICD-9-CM CODES FOR CANCERS TO BE STUDIED

All ICD-9-CM codes will be mapped to ICD-10-CM codes when the US data sources transition to the updated coding system.

Table 3-1: ICD-9-CM Codes for Female Breast Cancer

ICD-9-CM Codes	Description
174.0	Malignant neoplasm of nipple and areola of female breast
174.1	Malignant neoplasm of central portion of female breast
174.2	Malignant neoplasm of upper-inner quadrant of female breast
174.3	Malignant neoplasm of lower-inner quadrant of female breast
174.4	Malignant neoplasm of upper-outer quadrant of female breast
174.5	Malignant neoplasm of lower-outer quadrant of female breast
174.6	Malignant neoplasm of axillary tail of female breast
174.8	Malignant neoplasm of other specified sites of female breast
174.9	Malignant neoplasm of breast (female), unspecified

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-2: ICD-9-CM Codes for Bladder Cancer

ICD-9-CM Codes	Description
188.0	Malignant neoplasm of trigone of urinary bladder
188.1	Malignant neoplasm of dome of urinary bladder
188.2	Malignant neoplasm of lateral wall of urinary bladder
188.3	Malignant neoplasm of anterior wall of urinary bladder
188.4	Malignant neoplasm of posterior wall of urinary bladder
188.5	Malignant neoplasm of bladder neck
188.6	Malignant neoplasm of ureteric orifice
188.8	Malignant neoplasm of other specified sites of bladder
188.9	Malignant neoplasm of bladder, part unspecified
233.7	Carcinoma in situ of bladder

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-3: ICD-9-CM Codes for Colon/Rectum Cancer

ICD-9-CM Codes	Description
153.0	Malignant neoplasm of hepatic flexure
153.1	Malignant neoplasm of transverse colon
153.2	Malignant neoplasm of descending colon
153.3	Malignant neoplasm of sigmoid colon

ICD-9-CM Codes	Description
153.4	Malignant neoplasm of cecum
153.5	Malignant neoplasm of appendix vermiformis
153.6	Malignant neoplasm of ascending colon
153.7	Malignant neoplasm of splenic flexure
153.8	Malignant neoplasm of other specified sites of large intestine
153.9	Malignant neoplasm of colon, unspecified site
154.0	Malignant neoplasm of rectosigmoid junction
154.1	Malignant neoplasm of rectum
154.8	Malignant neoplasm of other sites of rectum, rectosigmoid junction, and anus
209.10	Malignant carcinoid tumor of the large intestine, unspecified portion
209.12	Malignant carcinoid tumor of the cecum
209.13	Malignant carcinoid tumor of the ascending colon
209.14	Malignant carcinoid tumor of the transverse colon
209.15	Malignant carcinoid tumor of the descending colon
209.16	Malignant carcinoid tumor of the sigmoid colon
209.17	Malignant carcinoid tumor of the rectum

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-4: ICD-9-CM Codes for Corpus Uteri

ICD-9-CM Codes	Description
179	Malignant neoplasm of uterus, part unspecified
182.0	Malignant neoplasm of corpus uteri, except isthmus
182.1	Malignant neoplasm of isthmus
182.8	Malignant neoplasm of other specified sites of body of uterus

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-5: ICD-9-CM Codes for Lung Cancer

ICD-9-CM Codes	Description
162.0	Malignant neoplasm of trachea
162.2	Malignant neoplasm of main bronchus
162.3	Malignant neoplasm of upper lobe, bronchus or lung
162.4	Malignant neoplasm of middle lobe, bronchus or lung
162.5	Malignant neoplasm of lower lobe, bronchus or lung
162.8	Malignant neoplasm of other parts of bronchus or lung
162.9	Malignant neoplasm of bronchus and lung, unspecified

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-6: ICD-9-CM Codes for Ovarian Cancer

ICD-9-CM Codes	Description
183.0	Malignant neoplasm of ovary
183.2	Malignant neoplasm of fallopian tube
183.3	Malignant neoplasm of broad ligament of uterus
183.8	Malignant neoplasm of other specified sites of uterine adnexa
183.9	Malignant neoplasm of uterine adnexa, unspecified site

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-7: ICD-9-CM Codes for Prostate Cancer

ICD-9-CM Codes	Description
185	Malignant neoplasm of prostate

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-8: ICD-9-CM Codes for Stomach Cancer

ICD-9-CM Codes	Description
151.0	Malignant neoplasm of cardia
151.1	Malignant neoplasm of pylorus
151.2	Malignant neoplasm of pyloric antrum
151.3	Malignant neoplasm of fundus of stomach
151.4	Malignant neoplasm of body of stomach
151.5	Malignant neoplasm of lesser curvature of stomach, unspecified
151.6	Malignant neoplasm of greater curvature of stomach, unspecified
151.8	Malignant neoplasm of other specified sites of stomach
151.9	Malignant neoplasm of stomach, unspecified site

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-9: ICD-9-CM Codes for Melanoma of Skin

ICD-9-CM Codes	Description
172.0	Malignant melanoma of skin of lip
172.1	Malignant melanoma of skin of eyelid, including canthus
172.2	Malignant melanoma of skin of ear and external auditory canal
172.3	Malignant melanoma of skin of other and unspecified parts of face
172.4	Malignant melanoma of skin of scalp and neck
172.5	Malignant melanoma of skin of trunk, except scrotum
172.6	Malignant melanoma of skin of upper limb, including shoulder
172.7	Malignant melanoma of skin of lower limb, including hip
172.8	Malignant melanoma of other specified sites of skin
172.9	Melanoma of skin, site unspecified

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

Table 3-10: ICD-9-CM Codes for Non-Hodgkin Lymphoma

ICD-9-CM Codes	Description
200.0	Reticulosarcoma
200.1	Lymphosarcoma
200.2	Burkitt’s tumor or lymphoma
200.3	Marginal zone lymphoma, unspecified site, extranodal and solid organ sites
200.4	Mantle cell lymphoma
200.5	Primary central nervous system lymphoma
200.6	Anaplastic large cell lymphoma
200.7	Large cell lymphoma
200.8	Other named variants of lymphosarcoma and reticulosarcoma
202.0	Nodular lymphoma
202.1	Mycosis fungoides
202.2	Sezary’s disease
202.7	Peripheral T-cell lymphoma
202.8	Other malignant lymphomas

ICD-9-CM = *International Classification of Diseases, 9th Edition, Clinical Modification*.

APPENDIX 4. READ CODES FOR HEALTH CARE UTILIZATION

Table 4-1: Read Codes for Mammography

Read code	Description
537…11	Mammography - X-ray
5372.00	Mammography normal
5373.00	Mammography abnormal
7P0F200	Mammography
5376.00	Mammography attended
6862.11	Mammography - screening

NOS = not otherwise specified.

Source: Medical and product dictionary browsers, version 3.0 London: General Practice Research Database (now the Clinical Practice Research Datalink); September 2015.

Table 4-2: Read Codes for Urine Cytology

Read code	Description
4KD…00	Urine cytology
4KD0.00	Urine cytology normal
4KD1.00	Urine cytology abnormal
4KD2.00	Urine cytology borderline
R119.00	[D]Abnorm find on cytological & histological exam of urine

Source: Medical and product dictionary browsers, version 3.0 London: General Practice Research Database (now the Clinical Practice Research Datalink); September 2015.

Table 4-3: Read Codes for Cystoscopy

Read code	Description
7B27.00	Endoscopic extirpation of bladder lesion
7B27.11	Cystoscopic extirpation of bladder lesion
7B27.12	Endoscopic removal of bladder lesion
7B27.13	TURBT - transurethral resection of bladder tumor
7B27000	Unspec cystoscopy and transurethral resection bladder lesion
7B27100	Unspecified cystoscopy and cystodiathermy
7B27200	Other unspecified cystoscopic destruction of bladder lesion
7B27300	Rigid cystoscopy and TUR bladder lesion
7B27400	Rigid cystoscopic diathermy of lesion of bladder
7B27411	Rigid cystoscopic cauterization of lesion of bladder
7B27500	Other rigid cystoscopic destruction of bladder lesion
7B27600	Flexible cystoscopic excision of bladder lesion

Read code	Description
7B27700	Flexible cystoscopy and cystodiathermy to bladder lesion
7B27711	Flexible cystoscopy and cauterization of bladder lesion
7B27800	Other flexible cystoscopic destruction of bladder lesion
7B27900	Endoscopic destruction of bladder tumor by laser
7B27y00	Other specified cystoscopic extirpation of bladder lesion
7B27z00	Cystoscopic extirpation of bladder lesion NOS
7B29.00	Other therapeutic cystoscopy
7B29.11	Other therapeutic endoscopic operations on bladder
7B29300	Endoscopic removal of blood clot from bladder
7B29311	Cystoscopic removal of blood clot from bladder
7B29y00	Other specified other therapeutic cystoscopy
7B29z00	Other therapeutic cystoscopy NOS
7B2A.00	Diagnostic cystoscopy
7B2A.11	Diagnostic endoscopic examination of bladder
7B2A000	Unspecified diagnostic cystoscopy & biopsy of bladder lesion
7B2A100	Unspec diagnostic cystoscopic exam bladder & biopsy prostate
7B2A200	Diagnostic cystoscopy using rigid instrument
7B2A300	Diagnostic cystoscopy & biopsy bladder lesion - rigid instr
7B2A400	Check cystoscopy using rigid instrument
7B2A500	Check cystoscopy and biopsy bladder lesion -rigid instrument
7B2A600	Diagnostic cystoscopy using flexible instrument
7B2A700	Diag cystoscopy & biopsy bladder lesion -flexible instrument
7BA800	Check cystoscopy using flexible instrument
7BA900	Check cystoscopy+biopsy bladder lesion - flexible instrument
7B2AA00	Diagnost endos exam bladder biop lesion bladder rigid cysto
7B2AB00	Diagnostic endoscop exam bladder biop lesion pros rigid cys
7B2AC00	Diagnostic endoscopic examination bladder using cystoscope
7B2AD00	Diag endoscop examination bladder biopsy lesion bladder NEC
7B2AE00	Diag endoscop examination bladder biopsy lesion prostate NEC
7B2Ay00	Other specified diagnostic cystoscopy
7B2Az00	Diagnostic cystoscopy NOS
7B2Az11	Check cystoscopy - unspecified

NEC = not elsewhere classified; NOS = not otherwise specified; OS = otherwise specified.

Source: Medical and product dictionary browsers, version 3.0 London: General Practice Research Database (now the Clinical Practice Research Datalink); September 2015.

APPENDIX 5. COVARIATES TO INCLUDE IN THE PROPENSITY SCORE MODEL

Additional variables that are risk factors for cancer are specified in Section 3.4.3, Table 1 through Table 3.

Demographic or Lifestyle	Indicators of Diabetes Severity
Age Sex Calendar year Body mass index > 30 or obesity surgery Smoking history History of alcohol abuse Socioeconomic status: Index of multiple socioeconomic deprivation, quintiles: first least deprived, fifth most deprived (CPRD) Duration of lookback period Calendar year Geographic region	Renal insufficiency or diabetic nephropathy Retinopathy Neuropathy Peripheral vascular disease Coronary heart disease Cerebrovascular disease Amputations Time since first diagnosis of type 2 diabetes mellitus, if available HbA1C (where available)
Medical Comorbidities	
Chronic disease score[a] Autoimmune disease Rheumatoid arthritis	Benign mammary dysplasia Urinary infections (chronic or recurring) Urinary cystitis Kidney stones Bladder stones Familial adenomatous polyposis Adenomatous colorectal polyps Crohn’s disease Ulcerative colitis Polycystic ovarian syndrome Benign prostatic hypertrophy Immunosuppressive diseases such as HIV/AIDS Peptic ulcer disease *Helicobacter pylori* infection
Medications	
Combined estrogen-progesterone hormone-replacement therapy Selective estrogen receptor modulators (raloxifene, tamoxifen) Cyclophosphamide Unopposed estrogen therapy Immunosuppressant (including oral and systemic corticoids) Opioids	

CPRD = Clinical Practice Research Datalink (UK); HIV = human immunodeficiency virus.

[a] For example, a score such as the Diabetes Complications Severity Index (DCSI) (Chang et al., 2012).

APPENDIX 6. ANTIDIABETIC DRUGS ELIGIBLE FOR INCLUSION IN THE COMPARATOR GROUP

Blood Glucose–Lowering Drugs (Excluding Insulin) by ATC Subgroup	Active Substance
A10BA, Biguanides	Metformin
A10BB, Sulfonamides, urea	Glibenclamide/glyburide
	Tolbutamide
	Gliclazide
	Glimepiride
	Carbutamide
	Chlorpropamide
	Tolazamide
	Glipizide
	Gliquidone
	Glyclopyramide
	Acetohexamide
A10BD, Combinations	Metformin/glibenclamide
	Metformin/rosiglitazone
	Rosiglitazone/glimepiride
	Pioglitazone/metformin hydrochloride
	Pioglitazone/glimepiride
	Sitagliptin/metformin hydrochloride
	Vildagliptin/metformin hydrochloride
	Pioglitazone/alogliptin
A10BF, Alpha glucosidase inhibitors	Acarbose
	Voglibose
	Miglitol
A10BG, Thiazolidinediones	Pioglitazone
A10BH, DPP-4 (dipeptidyl peptidase-4) inhibitors	Sitagliptin
	Vildagliptin
	Saxagliptin
	Linagliptin
	Alogliptin
A10BH, DPP-4 Combinations	Alogliptin/metformin
	Linagliptin/metformin

Blood Glucose–Lowering Drugs (Excluding Insulin) by ATC Subgroup	Active Substance
	Saxagliptin/metformin
A10BX, Other	Repaglinide
	Nateglinide
	Mitiglinide
	Exenatide
	Liraglutide
	Albiglutide
	Dulaglutide
	Lixisenatide

ATC = Anatomical Therapeutic Chemical (classification system).

Source: World Health Organization Collaborating Centre for Drug Statistics Methodology. ATC/DDD index 2015. Available at: http://www.whocc.no/atc_ddd_index/. Accessed 30 October 2015.