Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Cohort Profile: The Central Denmark Cancer Cohort

Emese Katalin Vágó, Henrik Toft Sørensen, Lars Pedersen, Erzsébet Horváth-Puhó

Department of Clinical Epidemiology, Center for Population Medicine, Aarhus University and Aarhus University Hospital, Aarhus, Denmark

Correspondence: Emese Katalin Vágó, Department of Clinical Epidemiology, Center for Population Medicine, Aarhus University and Aarhus University Hospital, Olof Palmes Allé 43-45, Aarhus N, 8200, Denmark, Email evago@clin.au.dk

Purpose: Cancer outcome studies frequently utilize registry data, which provide large-scale population-level information. However, these registries often lack detailed clinical information regarding comorbidities, lifestyle factors, and in-hospital treatments. The Central Denmark Cancer Cohort (CDCC) was established to address these limitations by linking medical and administrative registry data with clinical electronic medical records (EMR), to facilitate research to better understand the clinical course of cancer and its complications.

Patients and Methods: The CDCC includes all patients with incident cancer, except non-melanoma skin cancer, diagnosed in the Central Denmark Region between 2012 and 2021, with complete follow-up through December 31, 2021. The CDCC was identified from the Danish Cancer Registry and linked via the Civil Personal Registration number to the Central Denmark Region Clinical Information System (CDRCIS) and to national registries, including the Danish National Patient Registry, National Prescription Registry, and Register of Laboratory Results for Research. We extracted data on demographics, lifestyle factors, comorbidities, treatments, and survival outcomes and assessed the availability of these data.

Results: The CDCC included 68,028 patients with a median age of 68 years (interquartile range: 59–76 years); 47.2% were female. The most common cancers were prostate cancer (10,024 patients), breast cancer (9,100 patients), and non-small cell lung carcinoma (7,264 patients). At diagnosis, 98.0% of patients had laboratory test results available, 87.5% received at least one in-hospital medication, and 86.9% had at least one characteristic documented in CDRCIS. During median follow-up of 2.6 years, 37.8% of patients died and five-year survival was 59.0% (95% confidence interval: 58.6–59.4%). Data completeness varied by cancer type.

Conclusion: The CDCC integrates clinical, lifestyle, and laboratory data with cancer registry information. By capturing factors typically unavailable in registry-based research, this platform offers a unique foundation for longitudinal studies on the clinical course of cancer.

Keywords: electronic health record, cancer registry, real-world data, lifestyle factors, biomarkers, longitudinal study

Introduction

In 2022 there were over 18.7 million incident cases of cancer worldwide (excluding non-melanoma skin cancer) and over 48,000 cases in Denmark, according to GLOBOCAN 2022 estimates from the International Agency for Research on Cancer (IARC).[1] Cancer registries serve as the foundation for our current understanding of cancer outcomes.[2] These registries typically have information on the cancer disease, but not on comorbidities. Furthermore, registry data are often limited to treatment and stage at diagnosis, or shortly thereafter, and lack detailed clinical data. In registry-based cancer research, registry data therefore are often supplemented with data from other health and administrative registries. However, information on in-hospital treatments and lifestyle-related factors, such as alcohol consumption, smoking, or body mass index, is generally unavailable, making it difficult to examine fully the clinical course of cancer and its complications.

To address these limitations, our project aimed to merge data from the Central Denmark Region Clinical Information System (CDRCIS) and other Danish health registries into the Central Denmark Cancer Cohort (CDCC), to serve as a foundation for understanding cancer disease progression and complications. A distinguishing feature of this data

https://doi.org/10.2147/CLEP.S596863
Received: 21 January 2026
Accepted: 23 May 2026
Published: 8 June 2026

platform is the availability of lifestyle-related factors, such as smoking, alcohol consumption, and body mass index, recorded in electronic medical records, which are rarely available in registry-based cancer studies.

Materials and Methods

Data Sources

Denmark's tax-funded healthcare system provides free access to care at hospitals and general practitioners offices, along with partial reimbursement for most prescribed drugs. Health service utilization is extensively documented at the individual level in national health registries.[3,4]

Denmark's national health system has three administrative levels: state, regional, and municipal. Regions are responsible for hospitals and primary care services.[5] Denmark's public hospitals use two main electronic medical record (EMR) systems: the Columna Clinical Information System (CIS), developed by Systematic A/S, in the North Denmark Region, Central Denmark Region, and the Region of Southern Denmark; and Epic, developed by Epic Systems Corporation, in Region Zealand and the Capital Region of Denmark.[6] The Central Denmark Region (where Aarhus University is located) operates 10 non-psychiatric hospitals and serves approximately 1.3 million people, representing 23% of the total Danish population[7] (Figure 1).

The unique civil personal registration number (CPR number) assigned since 1968 to all Danish residents at birth or upon immigration makes it possible to link registries. The CDCC was constructed by using the following registries and data sources:

- The Danish Civil Registration System[8] (CRS) was established in 1968 as a comprehensive registry of demographic data for residents of Denmark. The System captures information on sex at birth, date of birth, date of death or emigration, and vital status. Moreover, it records parents' and spouses' CPR numbers, enabling linkage of family members.

Figure 1 Map of Denmark.
Notes: Light-shaded area: Remaining regions of Denmark. Boxed region: island of Bornholm.

- The Danish Cancer Registry[9] (DCR) has documented diagnoses of malignant neoplasms in Denmark since 1943, with mandatory reporting since 1987. Diagnoses have been coded using the *International Classification of Diseases, Tenth Revision* (ICD-10) since 1978, together with the diagnosis date. Tumor characteristics also are documented in the DCR, including morphology and topography (coded using ICD-O-3), TNM classification of stage, and Ann Arbor values for lymphomas. Data on cancer treatments administered within four months after diagnosis—including radiation, surgery, chemotherapy, and hormonal therapy—were registered in the DCR up until 2003, and since then in the Danish National Patient Registry.
- The Central Denmark Region Clinical Information System (CDRCIS) has collected data from medical records in the Central Denmark Region since 2012.[10] In addition to medications administered in hospitals, it includes data not recorded in nationwide health registries, such as patient characteristics, lifestyle-related data, imaging data, and surgical procedures.
- The Danish National Patient Registry[11] (DNPR) contains data on all Danish inpatient contacts since 1977. Data from emergency room and outpatient clinic visits were added starting in 1995. This Registry records information on each hospital contact, including admission and discharge dates, primary and secondary diagnoses, as well as records of examinations, surgeries, and treatments. Coding of diagnoses transitioned from ICD-8 to ICD-10 in 1994.
- The Danish Psychiatric Central Research Register[12] (DPCRR) has recorded all psychiatric hospital admissions in Denmark since 1969 and all outpatient and emergency contacts since 1995. In 1995, the register was integrated into the DNPR.
- The National Prescription Registry (NPR) contains data on prescriptions filled at community pharmacies in Denmark, with records available since 1995.[13] Primary variables include the dispensing date, the Anatomical Therapeutic Chemical (ATC) code, as well as package size and medication strength.
- The Register of Laboratory Results for Research (RLRR), established in 2008, includes routine biomarker test results retrieved from hospital laboratory information systems in Denmark.[14] The start date of data collection and specifics of variables captured varies among the Danish regions. Most tests are documented in accordance with the International System of Nomenclature, Properties, and Units (referred to as the NPU system),[15] with certain exceptions documented using national or regional codes.

Data availability in the registries linked to the CDCC is shown in Figure 2.

Cohort Construction

The CDCC was identified from the Danish Cancer Registry, focusing on persons with incident cancer diagnoses, excluding non-melanoma skin cancer, between 2012 and 2021. Using the CPR number, the data were linked to the CRS and restricted to persons who resided in the Central Denmark Region at the time of their incident cancer diagnosis, without any age restrictions. The cohort also has been linked to additional health and administrative registries, including the DNPR, DPCRR, NPR, and RLRR (Figure 2). Records from CDRCIS on patient characteristics and medications administered to hospitalized patients are also linked to the CDCC Cohort.

Statistical Analysis

Demographic information was retrieved from the CRS. Comorbidities present prior to the cancer diagnosis date were identified from the DNPR and the DPCRR and summarized using the Charlson Comorbidity Index score (CCI).[16] The CCI is a validated measure that assigns weighted scores to a set of selected chronic diseases to quantify an individual's overall comorbidity burden. The CCI was calculated using the original disease group definitions and weights,[16] mapped directly to ICD-8 and ICD-10 diagnosis codes (sTable 1). All available diagnosis history prior to the cancer diagnosis date was used to identify comorbidities (inpatient records from 1977 and outpatient records from 1995). To assess non-cancer comorbidities, we calculated a modified CCI score excluding the four cancer-related disease groups (non-metastatic solid tumor, leukaemia, lymphoma, and metastatic solid tumor). The prevalences of specific Charlson comorbidities, excluding cancers, also were calculated.

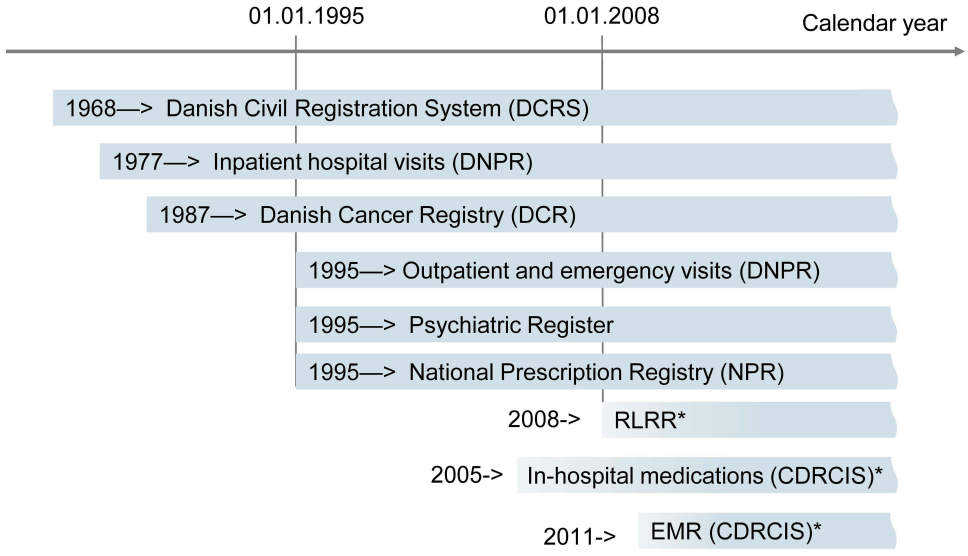

Figure 2 Data availability in the data sources linked to the CDCC.
Notes: *Not complete in the early years.
Abbreviations: RLRR, Register of Laboratory Results for Research; CDRCIS, Central Denmark Region Clinical Information System.

At the time of cancer diagnosis, laboratory test results (including hemoglobin level, platelet count, and leukocyte count) were obtained from the RLRR and patient characteristics (BMI, alcohol consumption, systolic and diastolic blood pressure, and smoking habits) were obtained from the CDRCIS. In this context, "at the time of cancer diagnosis" refers to the value of each characteristic closest to the date of diagnosis within the three months preceding it, or, if unavailable, within the following three months. Alcohol consumption is reported as a binary variable indicating whether consumption exceeded the sex-specific low-risk threshold (>7 units per week for women and >14 units per week for men), where one standard Danish drink unit is equivalent to 12 g of pure ethanol.[17]

We used data from the NPR to identify and report the ten most frequently dispensed medications, classified by their unique ATC codes, during the three-month period before or after the cancer diagnosis date.

For continuous variables, we calculated the median and interquartile range (IQR), and for categorical variables, we provided the proportion of patients. The proportion of patients with missing data also was calculated. Overall survival and corresponding 95% confidence intervals (CIs) were estimated using the Kaplan-Meier method.

Results

Cohort Characteristics

Between 2012 and 2021, the DCR documented a total of 314,249 incident cancer cases in Denmark, of which 68,028 (21.7%) were diagnosed in the Central Denmark Region. These patients were included in the CDCC. Patients were followed until December 31, 2021, death, or date of emigration, whichever occurred first. The median follow-up duration was 2.6 years, yielding a total of 226,794 person-years, during which 25,724 (37.8%) patients died. The five-year survival was 59.0% (95% CI: 58.6%, 59.4%).

Among the 68,028 patients in the CDCC, 98.0% had results for at least one of the selected laboratory tests recorded in the RLRR within three months before or after their cancer diagnosis date (Figure 3). Furthermore, 59,513 patients (87.5% of the CDCC) had medication data for at least one in-hospital stay registered in CDRCIS, and 59,143 (86.9%) had at least one patient characteristic (BMI, smoking status, alcohol consumption, or blood pressure measurements) documented in CDRCIS.

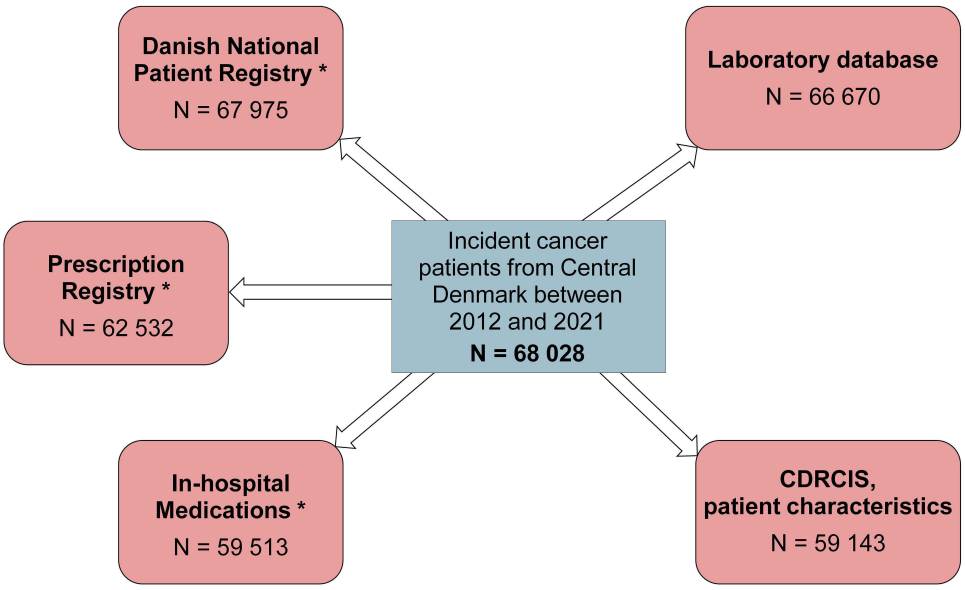

Figure 3 Selected data sources for the CDCC.
Notes: N: number of patients with at least one record within 90 days before or after cancer diagnosis. * Complete coverage. Patients not appearing in the database did not have a hospital visit, filled prescription or in-hospital medication.

Demographics and Comorbidities at Cancer Diagnosis

The median age in the CDCC was 68 years (IQR: 59–76 years) and 47.7% were female. Cancer diagnosis dates among CDCC members were evenly distributed across the study periods, with 39.2% occurring during 2012–2015, 30.4% during 2016–2018, and the remaining 30.4% during 2019–2021 (Table 1). Among all patients, 6.6% had moderate or severe comorbidity from conditions other than cancer, indicated by a CCI score (excluding cancer) of 3 or above. The

Table 1 Baseline Characteristics of the CDCC, Central Denmark Region, 2012–2021

	Calendar Period			Total CDCC
	2012–2015 (N = 26,675)	2016–2018 (N = 20,685)	2019–2021 (N = 20,668)	2012–2021 (N = 68,028)
Age				
Median (IQR)	68.0 (59.0–75.0)	68.0 (59.0–76.0)	69.0 (60.0–76.0)	68.0 (59.0–76.0)
Age categories (years)				
0–40	1263 (4.7%)	1060 (5.1%)	1096 (5.3%)	3419 (5.0%)
41–70	14,838 (55.6%)	10,766 (52.0%)	10,267 (49.7%)	35,871 (52.7%)
71=<	10,574 (39.6%)	8859 (42.8%)	9305 (45.0%)	28,738 (42.2%)
Sex				
Male	13,957 (52.3%)	11,061 (53.5%)	10,934 (52.9%)	35,952 (52.8%)
Female	12,718 (47.7%)	9624 (46.5%)	9734 (47.1%)	32,076 (47.2%)
Charlson Comorbidity Index score, excluding cancer				
0–1	22,853 (85.7%)	17,684 (85.5%)	17,553 (84.9%)	58,090 (85.4%)
2	2089 (7.8%)	1667 (8.1%)	1722 (8.3%)	5478 (8.0%)
≥3	1733 (6.5%)	1334 (6.4%)	1393 (6.7%)	4460 (6.6%)

(*Continued*)

https://doi.org/10.2147/CLEP.S596863

Table 1 (Continued).

	Calendar Period			Total CDCC
	2012–2015 (N = 26,675)	2016–2018 (N = 20,685)	2019–2021 (N = 20,668)	2012–2021 (N = 68,028)
Cancer types				
Esophageal	368 (1.4%)	276 (1.3%)	207 (1.0%)	851 (1.3%)
Stomach	398 (1.5%)	378 (1.8%)	438 (2.1%)	1214 (1.8%)
Colon	2414 (9.0%)	2027 (9.8%)	1671 (8.1%)	6112 (9.0%)
Rectal	1267 (4.7%)	910 (4.4%)	788 (3.8%)	2965 (4.4%)
Liver	280 (1.0%)	231 (1.1%)	243 (1.2%)	754 (1.1%)
Biliary	144 (0.5%)	107 (0.5%)	129 (0.6%)	380 (0.6%)
Pancreatic	616 (2.3%)	502 (2.4%)	553 (2.7%)	1671 (2.5%)
Non-small cell lung carcinoma	2827 (10.6%)	2190 (10.6%)	2247 (10.9%)	7264 (10.7%)
Small cell lung carcinoma	425 (1.6%)	294 (1.4%)	276 (1.3%)	995 (1.5%)
Melanoma	1467 (5.5%)	1240 (6.0%)	1592 (7.7%)	4299 (6.3%)
Breast	3631 (13.6%)	2690 (13.0%)	2779 (13.4%)	9100 (13.4%)
Cervical	260 (1.0%)	186 (0.9%)	175 (0.8%)	621 (0.9%)
Uterine	621 (2.3%)	398 (1.9%)	426 (2.1%)	1445 (2.1%)
Ovarian	433 (1.6%)	280 (1.4%)	255 (1.2%)	968 (1.4%)
Prostate	3901 (14.6%)	3054 (14.8%)	3069 (14.8%)	10,024 (14.7%)
Testicular	248 (0.9%)	220 (1.1%)	231 (1.1%)	699 (1.0%)
Kidney	708 (2.7%)	579 (2.8%)	635 (3.1%)	1922 (2.8%)
Bladder	637 (2.4%)	540 (2.6%)	486 (2.4%)	1663 (2.4%)
Brain	372 (1.4%)	319 (1.5%)	293 (1.4%)	984 (1.4%)
Hodgkin lymphoma	116 (0.4%)	108 (0.5%)	78 (0.4%)	302 (0.4%)
Non-Hodgkin lymphoma	886 (3.3%)	755 (3.6%)	757 (3.7%)	2398 (3.5%)
Multiple myeloma	302 (1.1%)	266 (1.3%)	285 (1.4%)	853 (1.3%)
Other	4354 (16.3%)	3135 (15.2%)	3055 (14.8%)	10,544 (15.5%)
Stage				
I	7334 (27.5%)	5941 (28.7%)	6370 (30.8%)	19,645 (28.9%)
II	2788 (10.5%)	2125 (10.3%)	1968 (9.5%)	6881 (10.1%)
III	2820 (10.6%)	2040 (9.9%)	2103 (10.2%)	6963 (10.2%)
IV	4239 (15.9%)	3252 (15.7%)	3416 (16.5%)	10,907 (16.0%)
Missing	5140 (19.3%)	4192 (20.3%)	3756 (18.2%)	13,088 (19.2%)
Not applicable[a]	4354 (16.3%)	3135 (15.2%)	3055 (14.8%)	10,544 (15.5%)

Notes: [a]Stage is not identified for the "Other" cancer category.

prevalence of individual Charlson comorbidities, along with history of additional comorbid conditions such as venous thromboembolism, alcohol and substance abuse, and schizophrenia, are described in sTable 2.

Cancer Type and Stage

We defined 22 cancer types, which accounted for cancers in 85% of the patient population. The most prevalent cancer types were prostate (14.7%), breast (13.4%), and non-small cell lung carcinomas (10.7%) (Table 1). Patients with cancers falling outside the 22 pre-defined types are described in sTable 3. Cancer stage, sex, and age distribution of the CDCC by cancer site are presented in sTable 4. Early diagnosis (stage I) was most frequent for prostate (62.0%), melanoma (56.0%), and uterine cancers (57.0%). Late-stage diagnosis (stage IV) predominated in pancreatic (48.1%), esophageal (42.2%), and stomach cancers (33.9%), consistent with their typical patterns of clinical presentation.

Data Completeness

Figures 4 and 5 illustrate the proportion of patients with specific biomarker assessments and the completeness of selected EMR data, stratified by baseline characteristics, cancer type and stage, and year of diagnosis. For members of the CDCC, the availability of biomarker data was high, while coverage for EMR elements was considerably lower. Among the EMR fields, the proportions of patients with recorded smoking status, BMI, and blood pressure were similar. In contrast, alcohol consumption was substantially less often. The most pronounced variation in data completeness occurred when stratifying by cancer type, particularly for biomarkers. More than half of cancer types had near 100% completeness for all three hematologic parameters, but others — especially melanoma, uterine, prostate, and testicular cancers —had lower coverage proportions.

Blood Pressure, BMI, Lifestyle Data, and Laboratory Results

Selected patient characteristics registered in CDRCIS at the time of cancer diagnosis are presented in Table 2. Smoking status was recorded for 43,032 patients; within this group, 36.9% were non-smokers, 37.5% were former smokers, 24.7% smoked daily, and 0.9% smoked occasionally. BMI at the time of cancer diagnosis fell within the normal weight category for 41.9% of Cohort members, within the overweight category for 35.1%, and within the obese category for 19.1%. Blood pressure measurements were available for 60.2% of Cohort members. Of the 25,219 patients with available alcohol consumption data, 3,545 (14.1%) reported consuming above the sex-specific low-risk threshold.

Laboratory test results (hemoglobin levels, thrombocyte and white blood cell counts) at the time of cancer diagnosis are presented in Table 2.

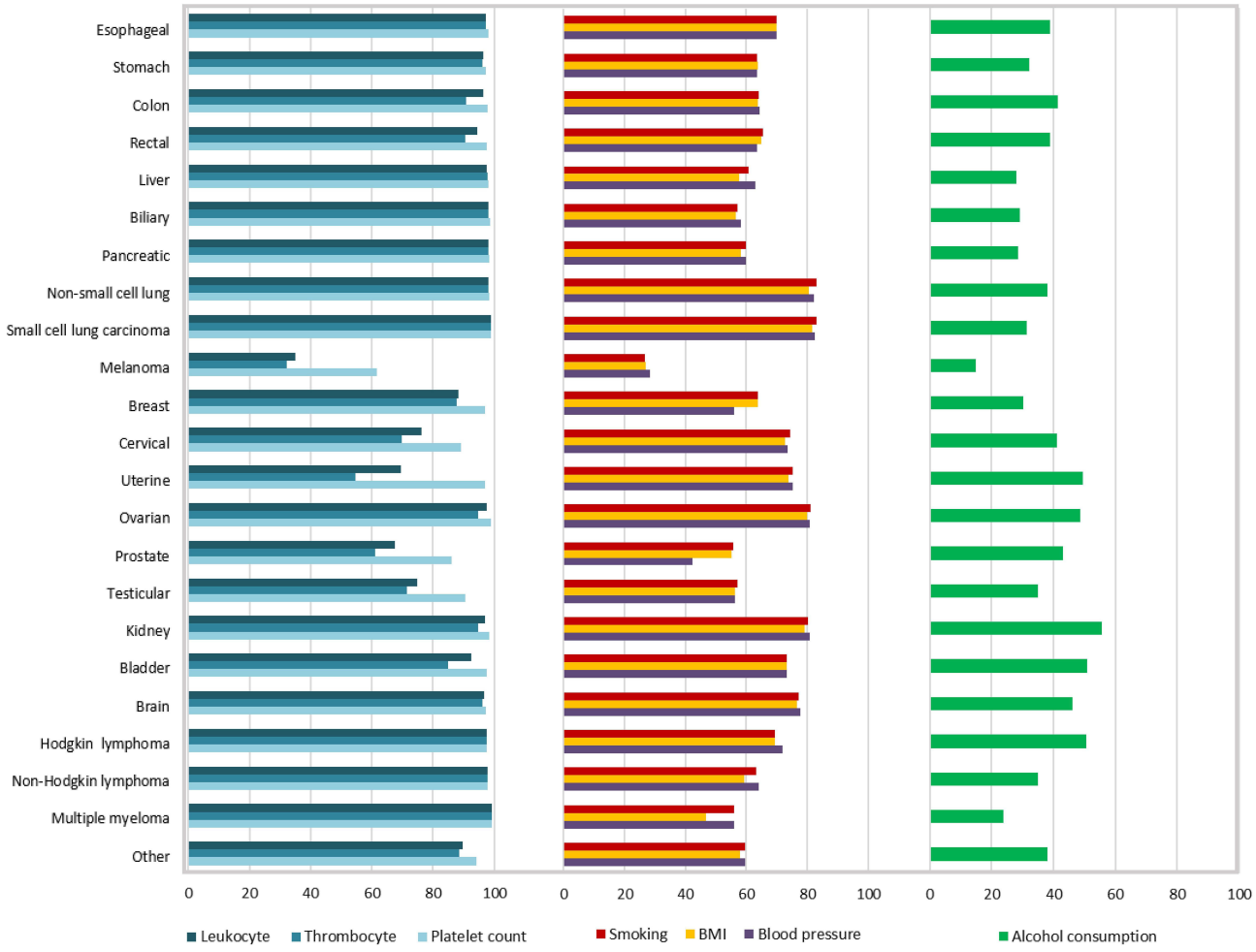

Figure 4 Proportion of patients with specific information recorded, by cancer type.

https://doi.org/10.2147/CLEP.S596863

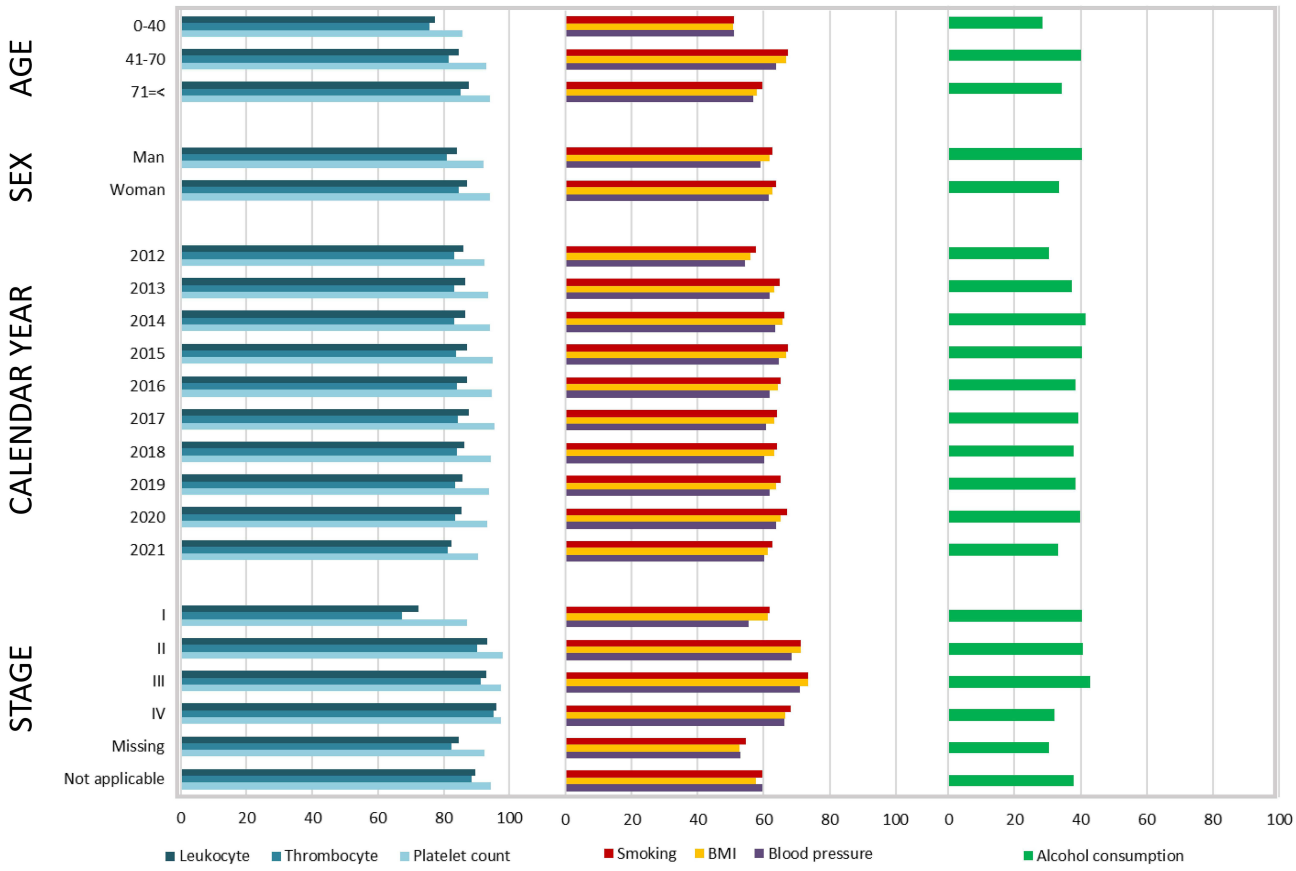

Figure 5 Ratio of patients with specific information recorded, by age, sex, calendar year, and cancer stage.

Cancer Treatments and Dispensed Medications

The primary cancer therapies during the three-month period before or after the cancer diagnosis date were identified using data from the DNPR, CDRCIS, and NPR (Table 2). Among CDCC members, 90.6% (61,639 people) received at least one form of cancer therapy. Surgery was the most common treatment for patients with solid tumors, administered to 73.4% of patients, followed by chemotherapy in 28.3% of patients and radiotherapy in 20.5%. Among patients with

Table 2 Selected Laboratory Results, Lifestyle Factors, and Cancer Treatments Among CDCC Members in the Three Months Before or After Cancer Diagnosis

Selected Biomarkers	
Hemoglobin (mmol/L)	
Median (IQR)	8.5 (7.6, 9.1)
Missing, N (%)	4726 (6.9%)
Thrombocyte (×10^9/L)	
Median (IQR)	271.0 (218.0, 339.0)
Missing, N (%)	11,798 (17.3%)
Leukocyte (×10^9/L)	
Median (IQR)	8.0 (6.4, 10.2)
Missing, N (%)	9935 (14.6%)

(*Continued*)

https://doi.org/10.2147/CLEP.S596863

Table 2 (Continued).

Clinical Measurements from the EMR	
Diastolic Blood Pressure	
Median (IQR)	79.0 (70.0, 89.0)
Missing, N (%)	27,061 (39.8%)
Systolic Blood Pressure	
Median (IQR)	138.0 (123.0, 155.0)
Missing, N (%)	27,053 (39.8%)
BMI	
Median (IQR)	25.4 (22.7, 28.7)
Missing, N (%)	25,752 (37.9%)
BMI, N (%)	
<18.5	1646 (3.9%)
18.5–24.9	17,730 (41.9%)
25.0–29.9	14,822 (35.1%)
30+	8078 (19.1%)
Missing	25,752 (37.9%)
Smoking, N (%)	
Never smoked	15,882 (36.9%)
Former smoker	16,116 (37.5%)
Current smoker	11,034 (25.6%)
Smokes daily	10,634 (24.7%)
Smokes occasionally	400 (0.9%)
Missing	24,996 (36.7%)
Alcohol consumption, N (%)	
<=7/14 units/week [a]	21,674 (85.9%)
>7/14 units/week [a]	3545 (14.1%)
Missing	42,809 (62.9%)
Selected Treatments [b]	
	Solid tumors [c] (N = 64,475) / **Non-solid tumors (N = 3553)**
Hormone therapy	7728 (12.0%) / 18 (~0.5%)
Surgery	47,325 (73.4%) / 2073 (58.3%)
Radiotherapy	13,227 (20.5%) / 323 (9.1%)
Chemotherapy	18,218 (28.3%) / 2232 (62.8%)
Targeted therapy	3227 (5.0%) / 2052 (57.8%)

Notes: [a] One standard drink unit was equivalent to 12 g of pure ethanol. The low-risk threshold during the study period was 7 units per week for women and 14 units per week for men. [b] Administered treatments registered in the Danish National Patient Registry, CDRCIS (in-hospital treatments), or filled prescriptions registered in the National Prescription Registry in the 3 months before or after cancer diagnosis. [c] Non-solid tumors: Hodgkin lymphoma, Non-Hodgkin lymphoma, and multiple myeloma.

hematological malignancies, targeted therapy (57.8%) and chemotherapy (62.8%) and surgery (58.3%) were the primary treatments.

The most frequently dispensed medications were for pain management, cardiovascular conditions, and gastrointestinal disorders (sTable 5). Paracetamol was the most commonly dispensed drug, used by 49.0% of patients. One-fifth of CDCC members redeemed prescriptions formorphine in outpatient care for managing severe pain. Pantoprazole, a proton pump inhibitor, was dispensed to 24.3% of patients. Cardiovascular medications, including simvastatin, metoprolol, and amlodipine, were also frequently dispensed (each to approximately 15% of Cohort members). This is consistent with the

observation that cardiovascular disease was among the most common comorbidities. Over the study period, a notable shift in statin-dispensing patterns was observed, with atorvastatin increasingly replacing simvastatin as the preferred lipid-lowering agent. Anti-inflammatory drugs, such as acetylsalicylic acid and ibuprofen, and antibiotics such as pivmecillinam were also frequently dispensed, with prescriptions redeemed by 17.3%, 15.6%, and 12.9% of Cohort members, respectively.

Discussion

We established a data platform called the Central Denmark Cancer Cohort, encompassing incident cancer patients diagnosed between 2012 and 2021 in Central Denmark. This platform allows for updates, enabling potentially unlimited follow-up in future epidemiological studies. The CDCC integrates clinical and demographic data from multiple national registries, including the DCR, DNPR, DPCRR, NPR, and RLRR. From CDRCIS, we extracted patient characteristics and in-hospital medication records.

A distinguishing feature of this data platform is its inclusion of detailed patient characteristics and lifestyle factors from CDRCIS. Specifically, we linked data to Cohort members on BMI, smoking, alcohol consumption, and blood pressure at the time of cancer diagnosis. Comprehensive laboratory test results are also available longitudinally for cohort members. To demonstrate the CDCC's potential for detailed biomarker analyses, we presented data on hemoglobin, platelet counts, and white blood cell counts at the time of cancer diagnosis.

The CDCC provides a resource for validating clinical risk scores and prediction models in cancer patients. It has been used already in a validation study by Lanting et al,[18] who evaluated a new cancer-associated thrombosis (CAT) risk score in over 12,000 cancer patients It was found to be superior in predicting 6-month VTE risk compared to the Khorana score. The results supported using the new CAT score to identify high-risk patients.

Strengths and Limitations

The CDCC's major strengths include its large sample size and the comprehensive integration of data from multiple sources, including detailed information on cancer characteristics but also on treatments, laboratory test results, comorbidities, and clinically relevant patient characteristics and lifestyle factors. Moreover, this information was available longitudinally throughout the 10-year study period, with the possibility of future updates.

This data platform also has several limitations concerning data completeness. The patient characteristics recorded in CDRCIS are missing for a substantial number of Cohort members, necessitating careful evaluation of patterns of missing data and the potential use of multiple imputation techniques in analyses. We also observed variation in biomarker availability across cancer types. This is likely attributable to differences in clinical guidelines and standard care protocols; routine complete blood count testing, which includes these hematologic parameters, may not be recommended to the same degree across all malignancies. EMR data completeness showed the greatest variability when stratified by cancer type, indicating that missing data are not missing completely at random. The particularly low data completeness observed in melanoma patients may be attributed to the potential outpatient management of this malignancy, where comprehensive lifestyle factor documentation might receive lower priority than in inpatient oncology settings.

The missingness mechanism likely differs across variables. For lifestyle factors such as alcohol consumption and smoking, missing not at random was plausible, as patients with higher consumption may be less likely to have these data recorded. For laboratory results, where testing follows standardised clinical protocols, missingness was more likely to be at random, driven by clinical setting and cancer type rather than test values themselves.

In this cohort profile study, we used a ±3-months time window around the cancer diagnosis date to identify baseline characteristics. In future studies using the CDCC, researchers may select different time windows depending on the research question and the defined index date. The lower data completeness observed for 2021 for smoking, BMI, blood pressure, and alcohol consumption might be due to a lag in EMR data processing, as these data were extracted in November 2021. This could be resolved in future database updates.

Perspectives

The CDCC offers a unique and rich data source for cancer research, especially in terms of in-hospital treatments and patient characteristics recorded in CDRCIS. The Cohort has the potential to be expanded both in size and in the scope of included data. Furthermore, the structured and coded nature of the underlying registries provides a foundation for transforming the data to a common data model, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM),[19] thereby enabling participation in international network studies.[20] Danish health registry data have already been successfully mapped to the OMOP CDM, as demonstrated by Sing et al[21] who analysed hip fracture epidemiology across 19 countries using OMOP-standardised patient-level data, including data from Denmark. The CDCC's integration of clinical and administrative health data aligns with the objectives of the European Health Data Space initiative,[22] which aims to promote the secondary use of health data for research across the European Union.

Detailed documentations of the Danish health and administrative registries linked to the CDCC are publicly available through the Danish Health Data Authority.[23] The use of international coding standards (ICD-8, ICD-10, ATC, NPU) supports interoperability. Although individual-level data cannot be transferred outside secure research environments in accordance with Danish data protection legislation, external researchers may initiate collaborative research by contacting the Department of Clinical Epidemiology, Aarhus University.

Currently, the CDCC is restricted to Central Denmark. As EMR systems are also implemented in the other four regions of Denmark, data from these regions could be incorporated into the Cohort in the future.

Imaging data are also recorded in CDRCIS and may be available for future research initiatives. Although pathology data are not currently included in the data platform, a national pathology registry[24] exists in Denmark, furthermore blood and tumor samples from patients are available in the Danish Cancer Biobank.[25] These additional data sources could and should be linked to the CDCC in the future.

Data Sharing Statement

Due to data security issues, no patient-level data for members of the CDCC can be made available to other researchers.

Ethics

According to Danish legislation, informed consent and approval from an ethics committee are not required for registry-based studies. Data handling procedures complied with Statistics Denmark's data confidentiality policy. The study was reported to the Danish Data Protection Agency (record no. 2016-051-000001/1819). Access to medical charts was granted by the Central Denmark Region.

Funding

This study was supported by the Independent Research Fund Denmark (grant number 3101-00102B). The Department of Clinical Epidemiology, Aarhus University, receives funding for other studies in the form of institutional research grants to (and administered by) Aarhus University. The Department of Clinical Epidemiology, Aarhus University, confirms that none of these studies have any relation to the present study.

Disclosure

Professor Henrik Sørensen reports paid evaluations for University of Oslo, the Norwegian Research Council, the Independent Research Fund Denmark, the European Research Council, and EpidStrategies. The authors report no other conflicts of interest in this work.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA*. 2024;74(3):229–263. doi:10.3322/caac.21834
2. Pukkala E, Engholm G, Højsgaard Schmidt LK, et al. Nordic Cancer Registries – an overview of their procedures and data comparability. *Acta Oncologica*. 2018;57(4):440–455. doi:10.1080/0284186X.2017.1407039
3. Thygesen LC, Ersbøll AK. Danish population-based registers for public health and health-related welfare research: introduction to the supplement. *Scand J Public Health*. 2011;39(7_suppl):8–10. doi:10.1177/1403494811409654

4. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *CLEP*. 2019;11:563–591. doi:10.2147/CLEP.S179083
5. Birk HO, Vrangbæk K, Krasnik A, et al. Denmark: Health System Summary 2024. European Observatory on Health Systems and Policies; 2024.
6. Rosenkrantz O, Wheler J, Westphal Thrane MC, Pedersen L, Sørensen HT. The Danish National Hospital Medication Register: a Resource for Pharmacoepidemiology. *Clin Epidemiol*. 2024;16:783–792. doi:10.2147/CLEP.S487838
7. About Central Denmark Region. Central Denmark Region. Available from: https://www.rm.dk/om-os/English/english/. Accessed May 25, 2025.
8. Pedersen CB. The Danish Civil Registration System. *Scand J Public Health*. 2011;39(7_suppl):22–25. doi:10.1177/1403494810387965
9. Gjerstorff ML. The Danish Cancer Registry. *Scand J Public Health*. 2011;39(7_suppl):42–45. doi:10.1177/1403494810393562
10. Health Innovation - Central Denmark Region. Central Denmark Region; 2023. Available from: https://www.rm.dk/om-os/English/english/health/health-innovation/. Accessed May 25, 2026.
11. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *CLEP*. 2015;449. doi:10.2147/CLEP.S91125
12. Mors O, Perto GP, Mortensen PB. The Danish Psychiatric Central Research Register. *Scand J Public Health*. 2011;39(7_suppl):54–57. doi:10.1177/1403494810395825
13. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H, Sørensen HT, Hallas J, Schmidt M. Data resource profile: the Danish National Prescription Registry. *Int J Epidemiol*. 2016;dyw213. doi:10.1093/ije/dyw213
14. Arendt JFH, Hansen AT, Ladefoged SA, Sørensen HT, Pedersen L, Adelborg K. Existing data sources in clinical epidemiology: laboratory information system databases in Denmark. *CLEP*. 2020;12:469–475. doi:10.2147/CLEP.S245060
15. Pontet F, Magdal Petersen U, et al; Joint Committee on Nomenclature, Properties and Units (C-SC-NPU) of the IFCC and IUPAC. Clinical laboratory sciences data transmission: the NPU coding system. *Stud Health Technol Inform*. 2009;150:265–269.
16. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373–383. doi:10.1016/0021-9681(87)90171-8
17. Holm AL, Veerman L, Cobiac L, Ekholm O, Diderichsen F. Cost-effectiveness of preventive interventions to reduce alcohol consumption in Denmark. *PLoS One*. 2014;9(2):e88041. doi:10.1371/journal.pone.0088041
18. Lanting V, Vágó E, Horváth-Puhó E, et al. Validation of clinical risk assessment scores for venous thromboembolism in patients with cancer: a population-based cohort study. *J Thromb Haemost*. 2025;23(2):600–609. doi:10.1016/j.jtha.2024.10.021
19. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012;19(1):54–60. doi:10.1136/amiajnl-2011-000376
20. Haber R, Webster-Clark M, Pratt N, et al. Core concepts in pharmacoepidemiology: multi-database distributed data networks. *Pharmacoepidemiol Drug Saf*. 2025;34(7):e70177. doi:10.1002/pds.70177
21. Sing CW, Lin TC, Bartholomew S, et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. *J Bone Miner Res*. 2023;38(8):1064–1075. doi:10.1002/jbmr.4821
22. European Health Data Space Regulation (EHDS) - Public Health; 2026. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en. Accessed April 10, 2026.
23. Nationale sundhedsregistre. Available from: https://sundhedsdatastyrelsen.dk/data-og-registre/nationale-sundhedsregistre. Accessed April 10, 2026.
24. Erichsen R, Lash TL, Hamilton-Dutoit SJ, Bjerregaard B, Vyberg M, Pedersen L. Existing data sources for clinical epidemiology: the Danish National Pathology Registry and Data Bank. *Clin Epidemiol*. 2010;2:51–56. doi:10.2147/clep.s9908
25. Laugesen K, Mengel-From J, Christensen K, et al. A review of major Danish Biobanks: advantages and possibilities of health research in Denmark. *Clin Epidemiol*. 2023;15:213–239. doi:10.2147/CLEP.S392416