

DATA SOURCES AND CREATION OF DATA FILES

Definitions

Included in the present study were all cancer cases diagnosed before 9 November 1980 in the population of the 1970 Population and Housing Census in Denmark. The Cancer Registry data have not been linked directly to the original 1970 census data, but have been linked to the Danish Occupational Mortality Register. The Mortality Register includes both the census data and data on deaths and emigrations in the census population during the 10-year period 9 November 1970—8 November 1980. The definitions that we have used in

our study consequently follow those of the Mortality Register (14, 15).

Population

The population included in the Mortality Register comprises persons who were registered in the 1970 census as Danish citizens on 9 November 1970 and who were registered in the Central Population Register. The Central Population Register holds information on all persons living in Denmark since 1 April 1968, and each

person is assigned a unique personal identification number (81). Census participants who had emigrated, had disappeared, or were dead before 9 November 1970, according to the Central Population Register, were excluded from the Mortality Register population. Only persons aged 15–74 years on the date of the census were included in the Mortality Register population (ie, persons born between 9 November 1895 and 8 November 1955). Details concerning the definitions of the population, deaths, and emigrations contained in the Mortality Register have been published elsewhere (14, 15).

Deaths

The Mortality Register includes deaths which occurred in the census population during the period 9 November 1970–8 November 1980. Deaths were registered from yearly death tapes, based on data in the Central Population Register and on death certificates.

Emigrations

The Mortality Register includes the registration of emigrations and disappearances in the census population during the period 9 November 1970–8 November 1980. This information comes from the Central Population Register. Only one emigration was registered per person, and reimmigrations were ignored.

Cancer cases

The Cancer Register is tumor-based (ie, each diagnosed tumor has a separate record). Tumors have been registered since 1943, and all tumor records for patients alive on 1 April 1968 should include a valid personal identification number. For each tumor the date of diagnosis is registered with year and month. The Cancer Registry asks for reports of the following malignant tumors: carcinoma, sarcoma, leukemia, Hodgkin's and non-Hodgkin's lymphoma, multiple myeloma, and mycosis fungoides. It also requests notification of the following tumor-like and benign lesions: papillomas of the lower urinary tract, benign tumors of the central nervous system and meninges, carcinoma in situ, and epithelial dysplasia of the cervix uteri (82). We have included all these tumor types in our study.

Selection of relevant Cancer Register records

The following criteria had to be met for the tumor records to be selected from the Cancer Register for linkage with the Mortality Register: (i) the person had to have been born between 9 November 1895 and 8 November 1955, (ii) the person had to have a valid personal identification number, and (iii) the diagnosis had to take place before December 1980.

The Cancer Register record includes information on the patient's date of death or emigration, if any. However, this information was not taken into considera-

tion in the selection of relevant tumor records as the date of death or emigration was already known from the data in the Mortality Register, and the registration of these data in the Mortality Register was assumed to be more reliable than the registration in the Cancer Register.

The selected tumor records needed some editing before linkage with the Mortality Register. First, cancer cases notified in 1943–1977 were coded according to a modified version of the seventh revision of the International Classification of Diseases (ICD-7) (83), whereas cancer cases notified in 1978–1980 were coded according to ICD-0 (84). The ICD-0 codes are translated to ICD-7 codes at registration, and ICD-7 codes were thus available for all notified tumors. However, for tumors of the bladder and skin the shift in classification implied a change in rules for the registration of multiple primaries. In order to obtain consistency in the registration of multiple primaries throughout the 10-year period, we recoded cancer cases of the bladder and skin diagnosed in 1978–1980. Second, a listing of multiple primaries in the same organ revealed several invalid double registrations which had to be deleted. Third, a check of the tumor records for invalid and inconsistent codes also revealed some errors.

Table 2 shows that a total of 298 507 tumor records were selected from the Cancer Register. Of these 778 were deleted due to invalid double registrations and 286 due to invalid or inconsistent codes; thus 297 443 tumor records remained for linkage with the Mortality Register. Some 1000 tumor records had only the year of diagnosis specified. These were precoded as of July, but the code was changed in the linkage procedure if it proved to be inconsistent with the date of death registered in the Mortality Register.

Linkage between the records of the cancer and mortality registers

To safeguard the anonymity of the cancer patients, all personal identification numbers were encrypted before the linkage (ie, transformed to unique identification numbers which can be traced back to the original personal identification numbers only by use of a secret key). The encrypted personal identification numbers were used in the matching.

The linkage of the records of the Mortality Register and those of the Cancer Register resulted in two data files, a population file containing all the census information needed for the analysis and a cancer file containing information on tumors for successfully matched records from the Cancer Register. Together, these two files constitute the Danish Occupational Cancer Register.

The linkage procedure was tested in a pilot linkage of a 1 % random sample of the records of the mortality and cancer registers. Inconsistencies between the

records of the two registers were scrutinized and compared with data in the Central Population Register and in municipality population registers. The purpose of this exercise was to establish a procedure with which to correct errors in the total data sets of records from the mortality and cancer registers. The inconsistencies found in the pilot linkage are listed in table 3.

Error type 1

The first type of error found consisted of the date of diagnosis in the Cancer Register being later than the date of death in the Mortality Register. In the pilot linkage two records with this error were found in the Cancer Register. The manual check showed the date of diagnosis to be incorrect in the Cancer Register in both cases, but no other errors were revealed in the records of the Cancer Register. We consequently decided that in the linkage procedure the date of diagnosis in the Cancer Register would automatically be redefined to the date of death in the Mortality Register in cases of errors of type 1.

Error type 2

In three instances of the pilot linkage the date of diagnosis in the Cancer Register was later than the date of emigration in the Mortality Register. The patients all turned out to be reimmigrants, and these tumors were consequently diagnosed outside the individual risk periods considered in the present study. Thus we decided that, in the linkage procedure, records from the Cancer Register with this type of error (type 2) would be omitted.

Error type 3

In the pilot linkage 17 of the records from the Cancer Register listed no date of death/emigration although such a date was given in the Mortality Register. The manual check revealed no further errors in these records, and we decided that, in the linkage procedure, the Mortality Register date of death would be accepted without further check.

Error type 4

Error type 4 occurred when the date of death/emigration in the Cancer Register was not the same as that in the Mortality Register. There were 10 records from the Cancer Register with this error in the pilot linkage. The manual check revealed that the records from the Cancer Register had incorrect personal identification numbers in five cases. In order to avoid linkage of records not belonging to the same person, we decided to list all errors of this type (type 4) for manual inspection and correction.

Error type 5

In 15 instances there was no equivalent record in the Mortality Register for a record from the Cancer Register, despite a date of death/emigration of later than 9 November 1970 in the Cancer Register. Data from the Central Population Register showed this inconsistency to be due to reimmigration in six cases and death before 9 November 1970 in one case. The remaining cases represented, however, either persons for whom the personal identification number had been changed or persons who had been mistakenly omitted from the 1970 census. We decided to list all these errors for manual inspection and correction.

Results of the check for errors

Table 4 shows the result of the linkage of the total data sets from the mortality and cancer registers. Out of the total of 297 443 records from the Cancer Register, 272 046 were matched to records in the Mortality Register and included in the cancer file, while 25 114 remained unmatched. Errors were registered in a total of 4036 cases. As has already been described, the 283 matched records with error type 2 (date of diagnosis later in the Cancer Register than the date of

Table 2. Selected tumor records from the Cancer Register.

	Number of records
Selected	298 507
Multiple primaries in same organ	
Checked	2 128
Deleted	778
Invalid and inconsistent codes	
Checked	488
Deleted	286
Records for linkage with Danish Occupational Mortality Register	297 443

Table 3. Result of the pilot linkage of a 1 % sample of the records of the Danish Occupational Mortality Register and the records of the Cancer Register.

Records	Number of records
Danish Occupational Mortality Register	35 539
Cancer Register	2 999
Population file	35 539
Cancer file	2 720
Unmatched records from the Cancer Register	279
Records from the Cancer Register with errors ^a	
Error type 1 ^b	2
Error type 2 ^b	3
Error type 3 ^b	17
Error type 4 ^b	10
Error type 5 ^b	15

^a One record may be affected by one or more types of errors.

^b See the text for an explanation of the type of error.

Table 4. Result of the linkage between the cancer and mortality registers.

Records	Number of records	
	Before corrections	After corrections
Danish Occupational Mortality Register	3 565 487	3 565 487
Cancer Register	297 443	297 359
Population file	3 565 487	3 565 487
Cancer file	272 046	271 996
Unmatched records from the Cancer Register	25 114	24 986
Records from the Cancer Register with errors ^a	4 036	3 637
Error type 1 ^b	122	—
Error type 2 ^b	283	377
Error type 3 ^b	2 285	2 820
Error type 4 ^b	590	—
Error type 5 ^b	1 071	820

^a One record may be affected by one or more types of errors.

^b See the text for an explanation of the type of error.

emigration in the Mortality Register) were not included in the cancer file. Only 1071 unmatched records from the Cancer Register had an error of type 5. This result indicated that the majority of the 25 114 unmatched records from the Cancer Register belonged to persons registered as dead or emigrated before 9 November 1970.

For the 4036 records from the Cancer Register with errors, the encrypted personal identification number was transformed back to the original personal identification number. The 4036 records represented a total of 3900 persons. In a check of the inconsistencies between the data from the two registers, all present and historical data from the Central Population Register were consulted for these 3900 persons. Five could not be identified in the Central Population Register.

In the total data set 122 records from the Cancer Register had error type 1; these were not checked any further. Error type 2 was found for 283 records from the Cancer Register, and 2285 records had error type 3. Neither of these errors were checked further. Error type 4 appeared for 590 records, and for these a manual check was made of the data in the Cancer Register, the date of death/emigration in the Mortality Register and all the data in the Central Population Register. For 1071 records, error type 5 was revealed. If, according to the Central Population Register, the persons had either died before the census date of 9 November 1970 or immigrated during the follow-up period of 9 November 1970—8 November 1980 no further check was made. This was the case for 63 records. For the remaining 1008 records, the data from the Cancer Register and all the data in the Central Population Register were checked manually.

Thus a total of 1603 records from the Cancer Register with errors were checked manually, including the five records which were untraced in the Central Population Register. These checks were based primarily on the listed present and historical data from the Central Population Register. In case these data were insuffi-

cient, the original Cancer Register notification form with the full name and address was found, and inquiries were made to municipality population registries or to the Department of Health, which keeps the National Death Index. A total of 292 inquiries were made.

The result of the manual check is shown in table 5. A total of 84 records from the Cancer Register were deleted, and the personal identification number was changed on 487. The date of death/emigration was changed on 19 records from the Mortality Register.

The result of the relinkage of the records from the cancer and mortality registers after the corrections is summarized in table 4. After corrections a total of 297 359 records of the Cancer Register remained, of which 271 996 were included in the cancer file relating to 256 674 persons. The total number of records from the Cancer Register with errors decreased from 4036 to 3637 after the corrections. The change in the total number of errors was smaller than the number of corrections made because a record corrected for one type of error in the first linkage may have fallen into one of the other error categories in the second linkage. If, for instance, a record with error type 5 in the first linkage belonged to a person who had his or her personal identification number changed after the census date, this person's personal identification number would have been corrected, as a result of the manual procedure, to the personal identification number at the time of the census. The second linkage may then show that this person emigrated before the date of diagnosis. In this case the change of the personal identification number probably took place at the time of reimmigration. This type of corrected record will show up with error type 2 in the second linkage.

Creation of data files for the tabulation of cancer incidence

As described in the preceding section, the linkage of records from the mortality and cancer registers resulted in two data files, the population file (including all 3 565 487 records of the Mortality Register) and the cancer file (including all 271 996 successfully matched records of the Cancer Register). The first problem to be faced with respect to the further editing of the data was whether one single data file would be sufficient for the tabulation of cancer incidence. In mortality analyses the same data file can be used for tabulating different causes of death, as there is only one possible event per person. In cancer incidence analyses, however, one has to take into account the fact that one person may develop more than one tumor. The possible sequences of events are illustrated in figure 1. If only one possible tumor for each person is considered in the analysis, part of the available information will be lost. The size of this loss of information was estimated for two different types of data files with the models that follow. In this estimation all information

in the cancer file was taken into account independently of whether it represented a malignant tumor or one of the registered precancerous lesions.

Model 1

Risk population: All persons in the population file, without cancer before 9 November 1970.

Tumors: First tumor at/after 9 November 1970.

Person-years at risk: From 9 November 1970 until date of tumor/death/emigration or 8 November 1980, whichever came first.

The model implies that both the total number of tumors and the person-years at risk are correct for calculating the overall cancer incidence. But the model implies a loss of information for calculating cancer incidence for specific diagnoses, as neither subsequent tumors in persons with tumors developed before 9 November 1970 nor multiple primaries in persons with

Table 5. Results of the manual check for error types 4 and 5.^a

	Type of correction			
	No correction	Personal identification number corrected in Cancer Register	Date of death corrected in the Danish Occupational Mortality Register	Record of Cancer Register deleted
Error 4				
Disagreement between the mortality and cancer registers as to date of death <1 month	79	—	—	—
Disagreement between the mortality and cancer registers as to date of death >1 month, Cancer Register date incorrect	198 ^b	—	—	—
Incorrect personal identification number in Cancer Register record; patient dead before personal identification numbers were introduced	—	—	—	64
Incorrect personal identification number in Cancer Register record; patient's correct personal identification number found	—	188	—	—
Incorrect date of death in the record of the Danish Occupational Mortality Register; correct date found	—	—	20 ^c	—
Cancer Registry notification form not found, computerized name incorrect	—	—	—	20
Cancer Registry notification form not found, computerized name correct	21	—	—	—
Error 5				
Personal identification number changed in the Central Population Register after the census date	—	286	—	—
Personal identification number changed in the Central Population Register before the census date	—	8	—	—
Patient immigrated after the census date	478	—	—	—
Patient emigrated before the census date	16	—	—	—
Patient missing in the census population of the Danish Occupational Mortality Register	217	—	—	—
Untraced	3	—	—	—
Untraced in the Central Population Register	—	5	—	—
Total	1012	487	20^c	84

^a Date of death to be read date of death/emigration.

^b Only the date of death/emigration in the Danish Occupational Mortality Register was used in the analysis.

^c Representing 19 records of the Danish Occupational Mortality Register.

INCIDENCE FOR A SPECIFIC CANCER DIAGNOSIS (eg, COLON CANCER)

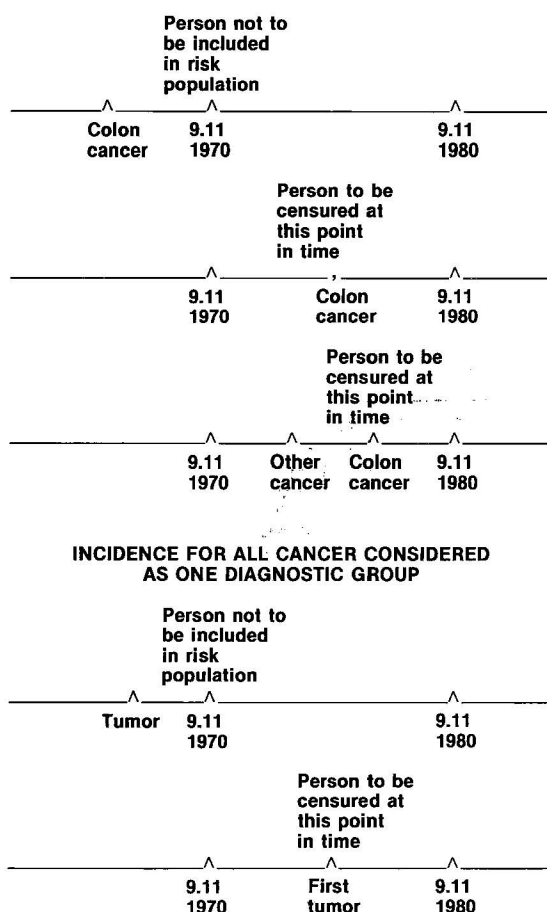


Figure 1. Possible sequences of events in tumor development related to the study period 9 November 1970–8 November 1980.

first tumor after 9 November 1970 are included. The size of the loss of information can be estimated based on the data in table 6 as:

$$\frac{G+H+I}{C+G+H+I} \times 100\% = \frac{2.3+0.2+2.6}{68.5+2.3+0.2+2.6} \times 100\% = 6.9\%$$

Model 2

Risk population: All persons in the population file.

Tumors: First tumor at/after 9 November 1970.

Person-years at risk: From 9 November 1970 until date of tumor/death/emigration or 8 November 1980, whichever came first.

This model has an advantage for calculating incidence rates for specific diagnoses in that persons who

developed a cancer before 9 November 1970 are included in the risk population after 9 November 1970. However, the model also implies two types of loss of information. First, multiple primaries diagnosed after 9 November 1970 are not included. Based on the figures in table 6, the size of this loss of information can be estimated as:

$$\frac{H+I}{C+G+H+I} \times 100\% = \frac{0.2+2.6}{68.5+2.3+0.2+2.6} \times 100\% = 3.8\%$$

Second, this model implies an overestimation of the person-years at risk used for calculating the incidence rates for specific diagnoses, since persons with a given diagnosis before 9 November 1970 will still be considered to be under risk for development of this specific diagnosis after 9 November 1970.

Person-years at risk for selected groups were calculated on the basis of a preliminary data file including all persons and all first tumors in an attempt to quantify the size of the overestimation in the person-years at risk. For persons with first tumor before 9 November 1970, the person-years at risk were counted from 9 November 1970 up until death/emigration or 8 November 1980. For persons without the first tumor before 9 November 1970, the person-years at risk were counted from 9 November 1970 up until first tumor/death/emigration or 8 November 1980. Out of a total of 33 219 525 person-years at risk, persons who developed lung cancer, colon cancer, and nonmelanoma skin cancer before 9 November 1970 only contributed with 7378, 20 763, and 77 270 person-years at risk, respectively. The overestimation of the person-years at risk is thus of minor significance with this model.

In order to avoid the errors described for models 1 and 2, we decided to create one data file for each of the diagnostic groups using model 3.

Model 3

Risk population: All persons without a tumor from the given diagnostic group diagnosed before 9 November 1970 in the population file.

Tumors: First tumor from the given diagnostic group at/after 9 November 1970.

Person-years at risk: From 9 November 1970 until date of tumor/death/emigration or 8 November 1980, whichever came first.

While this model is computationally more complicated, it makes full use of the data, and it is conceptually more satisfying. When trying to establish whether there is an association between a given occupation and the risk of a specific cancer, we did

not want the measurement to be influenced by whether or not the patients developed some other cancer before.

With the use of this set of data files the number of tumors and the person-years at risk will be "correct" for calculating both the overall cancer incidence and the incidence for specific diagnostic groups. Therefore the following five different types of persons were considered in the analysis: (i) persons who already had a given cancer on the census date, (ii) persons who died without the given cancer, (iii) persons who emigrated without the given cancer, (iv) persons who developed the given cancer, and (v) persons who were still present in the country and had not developed the given cancer at the end of the 10-year follow-up period.

Variables in the data set

The data content of the linked register can be obtained from the Danish Cancer Registry upon request.

Census information

In the census, information was collected from self-administered questionnaires which were checked by the municipalities and coded by Danmarks Statistik. Information was collected on economic status, industry, occupation, municipality of home, municipality of workplace, kind of dwelling, dwelling facilities, marital status, and household composition. The personal identification number gives the information on the date of birth and sex.

In the classification on economic status the population was divided into persons who were economically active and persons who were economically inactive at the time of the census. (A copy of this classification can be obtained from the Danish Cancer Registry upon request.)

Occupation and industry were recorded only for persons who were economically active. Occupation was coded according to special Danish nomenclature with which a distinction was made between the self-employed persons, family workers, salaried employees, skilled workers, and unskilled workers — a total of 218 codes were possible. (A copy of the detailed classification can be obtained from the Danish Cancer Registry upon request.)

Industry was coded according to a modified version of the International Standard Industrial Classification (ISIC), covering 245 codes. (A copy of the detailed classification can be obtained from the Danish Cancer Registry upon request.)

The combined codes for occupation and industry allows the analysis to be made for 53 410 ($= 218 \times 245$) specific subgroups. This number did not, of course, make sense in a population of approximately 2 mil-

lion economically active persons, and the combined codes for economic activity, occupation, and industry were therefore grouped.

First, a socioeconomic classification of 32 groups was made. A detailed description of this classification has been published elsewhere (14, 15), and a copy of the classification can be obtained from the Danish Cancer Registry upon request.

Second, a detailed occupational classification was constructed from the combination of the occupational and industrial codes. The purpose was to form homogeneous groups. One occupational code was, for example, "436 driver, coachman." This group was divided through the incorporation of the industrial codes (eg, driver in "400—409 construction," "702 trainway and omnibus," "710 taxi," etc). A total of 492 detailed occupational groups were formed for the men and 447 for the women. (A copy of the detailed occupational classification can be obtained from the Danish Cancer Registry upon request.)

Cancer Register information

The information of the Cancer Register comes from notification forms from hospital departments and private specialists in dermatology and gynecology, and from death certificates. Notification to the Cancer Register has, throughout the study period, been voluntary, but record-linkage studies (85, 86) indicate that about 95 % of diagnosed cancer cases have been registered. Information has been collected on the identity of the patient, date of diagnosis, diagnosis, extent of tumor,

Table 6. Distribution of first and subsequent tumors in relation to the census date 9 November 1970.

		Total (%)
All first tumors	A	94.4
First tumor before 9 November 1970	B	25.9
First tumor after 8 November 1970	C	68.5
All subsequent tumors	D	5.6
Subsequent tumors for persons with first tumor before 9 November 1970	E	3.1
Subsequent tumor before 9 November 1970	F	0.7
Subsequent tumor after 8 November 1970 and tumor number = 2	G	2.3
Subsequent tumor after 8 November 1970 and tumor number > 2	H	0.2
Subsequent tumor for persons with first tumor after 8 November 1970	I	2.6
Total		100.0 ^a

^a Number of tumors = 271 996.

basis for diagnosis, and treatment during the four months after diagnosis.

The diagnosis was coded according to a modified version of ICD-7 for the period 1943–1977 and according to ICD-0 for 1978–1980, as has already been described. The modified ICD-7 classification, which was available for all the tumor records, included 639 codes. In the analysis these were grouped into 47 diagnostic groups. (A copy of the detailed occupational classification can be obtained from the Danish Cancer Registry upon request.)

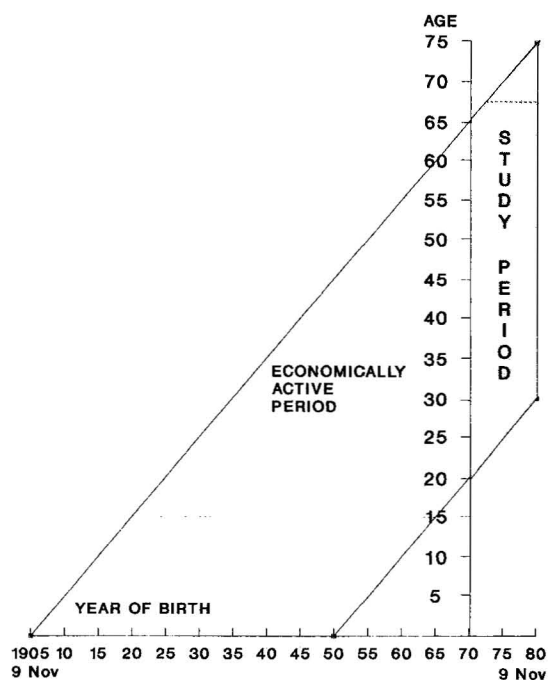


Figure 2. Study population defined by year of birth, economically active period, and study period.

Calculation of observed and expected numbers

After the diagnoses were grouped, 48 data sets were formed for the analysis according to model 3, as has already been described. There was one data set for each diagnostic group and one for the overall cancer incidence.

The data were analyzed separately for the men and the women, and only persons aged 20–64 years at the time of the census were included. See figure 2. The observed number of cancer cases and the person-years at risk were tabulated by five-year age groups defined by age at the time of the census. The cancer incidence rates for all economically active persons were used for calculating the expected number of cancer cases in a given social or occupational subgroup. We calculated the standardized incidence ratios by dividing the observed number of cancer cases by the expected number. For abbreviation the notation “relative risk” has been used in the text for this index. Ninety-five percent two-tailed confidence intervals were calculated for the relative risks on the basis of the assumption that the observed number of cases followed a Poisson distribution (87) and that, for observed numbers over 30, the distribution was normal (88).

Number of persons and cancer cases in the study

A total of 2 804 190 persons aged 20–64 years at the time of the census were included in the study. Of these persons, 38 597 had already been diagnosed with cancer before the census date; thus 2 765 593 persons were left at risk when all malignant neoplasms were considered as one diagnostic group. During the 10-year follow-up period 109 060 persons were diagnosed with at least one cancer. A total of 115 022 cancer cases were found when the cases were aggregated across the 47 specific diagnostic groups. See table 7.

Table 7. Number of persons and cancer cases.

	Men		Women	
	All men	All economically active men	All women	All economically active women
All persons	1 401 967	1 292 337	1 402 223	730 547
Persons at risk	1 389 232	1 281 874	1 376 361	720 450
Persons who developed cancer	52 853	47 538	56 207	25 557
Incident cancer cases	55 606	49 926	59 416	26 855