Using record linkage to construct a population-based cancer patient cohort with multiple disease outcomes

PART I

Methods and experiences of the Netherlands Cancer Institute in the construction of a population-based breast cancer survivor cohort linked with cardiovascular disease and mortality registries.

Naomi Boekel, MSc

Michael Schaapveld, PhD

Flora E van Leeuwen, PhD

Division of Psychosocial Research and Epidemiology

Netherlands Cancer Institute, 2015.

Introduction

Research on late adverse effects of cancer treatment is increasingly important in view of improved cancer survival. For our current study we wanted to assess the risk of cardiovascular disease incidence and mortality after radiotherapy and/or chemotherapy for breast cancer. For this study we constructed a large cohort of breast cancer survivors with detailed treatment information and to subsequently collect data on both cardiovascular morbidity and mortality after breast cancer. The data collection for this cohort was performed solely by linkages of different population-based registries and electronically kept data files. This was a complex process that involved many different parties and regulations to secure the privacy of patients. In the current paper, we describe the linkage procedures that were used to construct this cohort.

Box 1. The different parties involved in the construction of the first cohort.

The Netherlands Cancer Registry *(NCR; Integraal Kankercentrum Nederland, de Nederlandse Kankerregistratie: http://cijfersoverkanker.nl/gegevens-aanvragen-16.html):* The NCR is a population-based nationwide cancer registry and has a coverage of at least 96% of invasive malignant neoplasms and selected non-invasive cancers occurring in the Netherlands since 1989 [1].

The Steering Committee Heart Interventions Netherlands *(BHN, Begeleidingscommissie Hartinterventies Nederland: https://www.bhn-registratie.nl/)*: The BHN registers heart interventions (including open heart surgery and percutaneous coronary interventions) performed in the Netherlands since 1995. The heart interventions registry has complete coverage of cardiothoracic procedures in the Netherlands since 2000 [2].

Dutch Hospital Data *(DHD, Landelijke Medische Registratie (LMR)):* All hospital discharges in the Netherlands are registered by the DHD and data is electronically available for record linkage at Statistics Netherlands since 1995. Completeness of the LMR registry has been gradually decreasing since 2004 as a result of the introduction of a new system for financing health care in the Netherlands. The introduction of diagnosis-treatment combinations made additional recording of hospital discharge diagnosis less important for individual hospitals, which prompted hospitals to stop participation in the LMR registry. While only about 1% of all discharge diagnoses of the Dutch hospitals were not recorded in the LMR in 2004, this proportion had increased to almost 15% in 2009.

Statistics Netherlands (*CBS, Centraal Bureau voor de Statistiek: http://www.cbs.nl/nl-NL/menu/informatie/beleid/zelf-onderzoeken/default.htm):* The CBS keeps record of many different statistics (so called microdata), including the nation-wide cause of death registry. Individual researchers can get access to these microdata, either at the bureau of Statistics Netherlands or via remote access to Statistics Netherlands microdata, under strict conditions, when the director-general of statistics has authorized the institution the researcher is affiliated with to use the microdata facilities. Data from the DHD are available through CBS for diagnosis years 1995 and later.

Radiotherapy institutes: Many of the Dutch radiotherapy institutes have electronically registered details of patients' treatments.

European Organisation for Research and Treatment of Cancer *(EORTC):* The EORTC is an international organization that, among others, keeps data from randomised controlled trials.

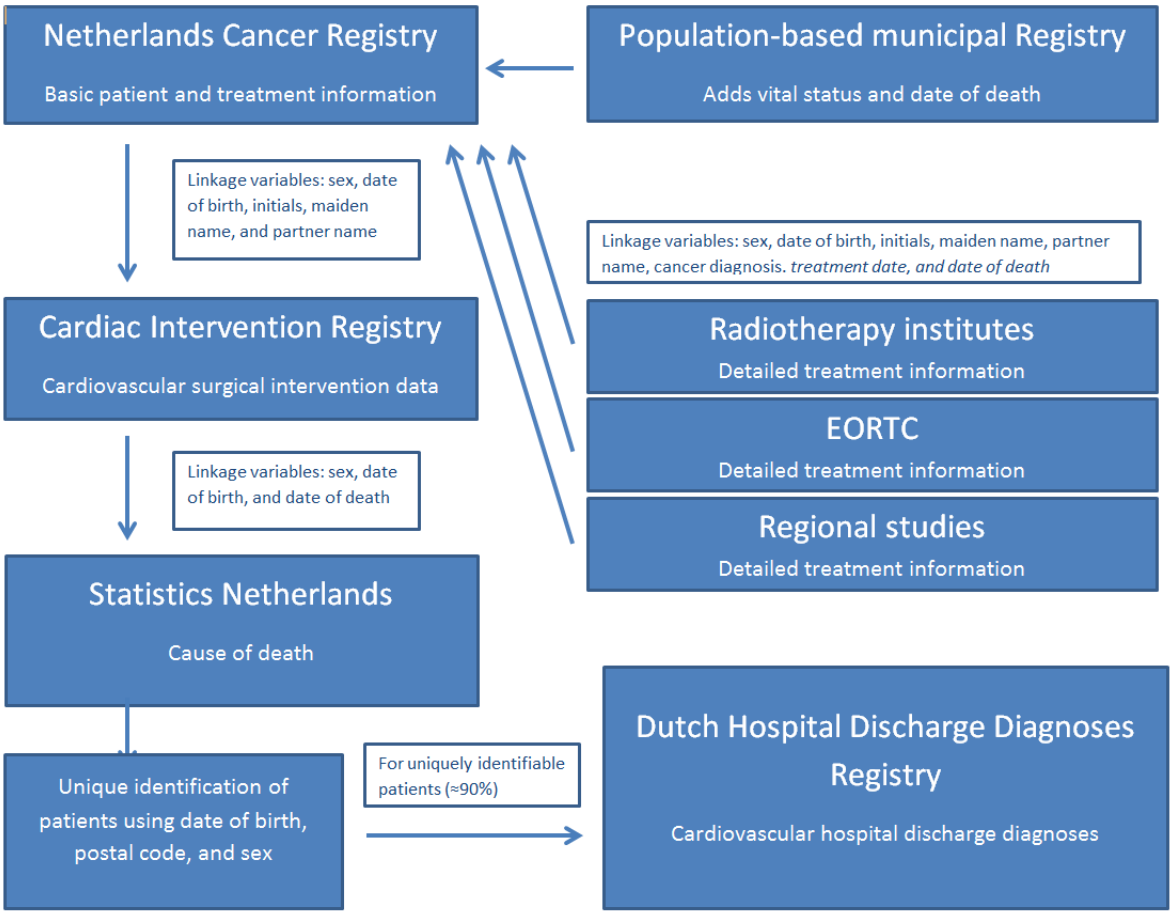
Regional study coordinators: Many regional studies have been performed in breast cancer survivors in the Netherlands; several have collected treatment details.

Methods

Data collection

The route of data collection and record linkage is shown schematically in Figure 1. The purpose of the subsequent linkages was to enrich a cancer patient cohort derived from the NCR with with information on cardiovascular disease incidence and mortality, as recorded by the BHN, CBS, and DHD.

Figure 1. Data collection route for establishing a population-based breast cancer cohort with cardiovascular disease registries



Linkage with and analysis of causes of death and hospital discharge diagnoses required a dedicated computer with remote access to Statistics Netherlands. This also meant that to us as researchers, the various linkage procedures between registries would be more or less a black box, with an anonymised data file as the end product.

The initial patient selection to construct a population-based breast cancer cohort was performed by the NCR. From their registry, two patient cohorts were selected: (1) all women diagnosed with ductal carcinoma in situ of the breast (DCIS) as their first neoplasm between 1989 and 2005 in the Netherlands, and (2) all women diagnosed with invasive breast cancer as their first tumour between 1989 and 2005 in

one of four former NCR regions (Comprehensive Cancer Centers Amsterdam, North, East, and South). In total, the cohorts consisted of respectively 12,309 and 93,630 patients. For both cohorts, the NCR provided the NCR-registration number, vital status, date of death, registered postal codes. For initial and subsequent neoplasia the date of diagnosis, topography, morphology, differentiation, stage, type of surgery, chemotherapy (yes/no), and radiotherapy (yes/no) were provided. The NCR acquires vital status and date of death through an annual linkage with the population-based municipal personal records database. The patients' identifying data (gender, maiden name, partner's name, date of birth) were encrypted at the NCR office using the encryption module of the heart intervention registry provided by KIK (department of Medical Informatics/Clinical Registry Office (Klinische InformatieKunde), Amsterdam University Medical Center). KIK maintains the heart intervention registry, provides the ICT environment and the technical support for the BHN heart intervention and acted as a trusted third party in our linkage process. After data encryption the data were sent to the heart intervention registry (KIK).

The KIK performed the linkage of the NCR data within the BHN registry, using date of birth, sex, and the first four letters of the maiden name as linkage variables. Additionally, the partner's last name as registered by the NCR was compared to the first four letters of the partner's name registered by BHN. Because of possible partnership changes, this variable was not used as a linkage variable. The KIK added the BHN registration number to the data, removed the NCR registration number and saved a linkage table containing both registration numbers for future additional linkages. Data on cardiovascular interventions were added to the cohort KIK uploaded the data to Statistics Netherlands for subsequent linkage with microdata.

Statistics Netherlands linked the causes of death to the uploaded dataset, using the unique combination of date of birth, date of death and sex. Subsequent linkage with DHD data provided hospital discharge diagnosis of cardiovascular diseases (CVD). DHD has recorded all (cardiovascular) hospital admissions, regardless of any surgical interventions performed, and thereby complemented the heart intervention registry. Furthermore, the DHD data also provided surgical interventions for CVD for the years in which the heart intervention registry did not yet fully cover the Netherlands.

Linkage with microdata, including the DHD, required the use of the RIN number, which is an encrypted version of the national identification number (Burgerservicenummer, BSN) that uniquely identifies each Dutch inhabitant: Statistics Netherlands does not use personal identifiers such as names for linkage. In order to be able to link our cohort to the DHD, the population register was used to determine whether the patients in our cohort were uniquely identifiable in the Netherlands at any moment in time between 1995 and the end of follow-up (death or January 1st 2010). The following variables were available in our cohort for unique identification: date of birth, sex and the first four digits of postal code. This meant that if a patient, somewhere between 1995 and end of follow-up, at any time was the only female living in a certain (four digit) postal code with the specific date of birth, she could be uniquely identified and her RIN-number was assigned to her. If a person was not uniquely identifiable using these three variables, full postal code, date of death and date of last vital status were used to try to uniquely identify the individual.

The biggest drawback of this method is that not all patients are uniquely identifiable. In our cohorts, 90.7% of the DCIS and 89.1% of the invasive breast cancer patients could be uniquely identified and matched to a single RIN number. In addition to the RIN number, Statistics Netherlands provided us with files including dates between which each person in the Netherland could or could not be uniquely identified based on sex, date of birth and a four digit postal code.

After these two linkages were performed, Statistics Netherlands removed all identifying variables (including date of birth, and postal codes) and encrypted the BHN registration number. Together with the DHD data which comprised all hospital discharge records for the whole of the Netherlands and the RIN number, data was made available to us via remote access: a stand-alone working station in our institution with software and hardware -finger print scanner- installed by Statistics Netherlands allowing remote access to microdata at Statistics Netherlands. We performed a linkage between our data and the DHD data using this RIN number.

Detailed treatment information

In addition to the limited treatment information that was provided by the NCR, we collected detailed treatment information from electronically kept files for part of the cohort. This information was gathered from radiotherapy institutes, clinical trials (through the European Organization on Research and Treatment for Cancer (EORTC, Brussels) or through trial coordinators), and regional studies. This included the following treatment information: date of treatment, radiation fields, total radiation dose, radiation dose per fraction, radiation beam energy, and chemotherapeutics for all female breast cancer patients registered in their database between 1989 and 2005. Not all variables were available from each institute/study, however. Furthermore, we requested identifying variables for linkage, including: maiden name, partner's name, date of birth and date of death. Unfortunately, the EORTC was not able to provide names.

As a privacy-preserving measure, all additional treatment data were sent to the NCR directly by all the institutes/study coordinators. The NCR linked these databases to the previously selected NCR cohort based on maiden name, partner's name, date of birth, sex, and cancer diagnosis. When more than one match was found using this method, treatment date and date of death were used to try to rule out all but one. In total, 93% of the 46,818 treatment records (including duplicates) could be allocated to one person in the selected NCR cohort. For the EORTC data, this percentage was lower (81%), due to the fact that EORTC did not provide patient names for record linkage with the NCR.

Long versus wide data structures and the selection of records

The NCR provided the data in a wide data structure, with only one record per person. The other registries, however, used long data structures with multiple records per person.

The BHN registry is divided into a registry for cardiology (comprising 30 cardiology centres; information on percutaneous coronary interventions (PCI) and transcatheter heart-valve interventions (THI)) and a registry for cardio-surgery (comprising 16 cardiothoracic surgery centres), and contains a new record for each cardiovascular intervention that was performed. The KIK added the "long" BHN data to the existing NCR data table, leading to multiple records per person, with the NCR data for a person being simply duplicated. For the two components of the registries, the KIK made a separate data file. For our survival analysis, we were interested in the first cardiovascular event that occurred after breast cancer diagnosis. Because cardiovascular event information in the first five years after breast cancer diagnosis was not available for the entire cohort (due to lack of digitalized CVD data before 1995), we were interested in the first cardiovascular event that occurred at least five years after breast cancer diagnosis. Therefore, for each type of cardiovascular intervention the first occurrence at least five years after breast cancer diagnosis was selected from the BHN data, and a wide data structure – with only one record per person – was created.

The DHD database also has a long data structure, with multiple records for each hospital discharge (see appendix). Linkage with this registry is, however, complicated by the fact that linkage is only possible for

periods during which a person was uniquely identifiable. To be able to select, for a certain cardiovascular disease diagnosis, the first hospital discharge (five years) after breast cancer diagnosis, we first had to identify the first period during which a person was uniquely identifiable (five years) after breast cancer diagnosis. This was done using the Statistics Netherlands "uniqueness" data files. After this period was ascertained, we selected the first hospital discharge per cardiovascular disease diagnosis during this "unique" period, and added this to the data file. During each hospital discharge, both a main diagnosis -giving the most important reason for the hospital admission- and one or multiple sub-diagnoses – giving information on other diseases that were diagnosed or complications that occurred during the admission – are registered. Because the CVD sub-diagnoses in patients with a main diagnosis other than CVD mostly seemed to be complications from other diseases, we chose to take into account only main diagnoses.

Performed data checks

Importance of partner's name

To check the usability of the partner's last name as a linkage variable, the KIK grouped the cohort into four mutually exclusive groups: (1) partner's names in NCR and BHN match, (2) NCR's partner's name is empty, BHN's partner's name is filled, (3) NCR's partner's name is filled, BHN's partner's name is empty, and (4) both partner's names are filled, but do not match. All of these patients had undergone a cardiovascular intervention – for otherwise they would not be in the BHN registry – and a cardiovascular intervention always entails a hospital admission. This made it possible to compare the percentage of patients who had a cardiovascular disease diagnosis recorded in the DHD between these four groups (table 1).

Table 1.

Linkage group	Total (n)	LMR match[a] (%)	Possible LMR match[b] (%)	No LMR match[c] (%)
Total	116	70	8	22
1	76	89	4	7
2	5	0	40	60
3	20	70	0	30
4	15	0	27	73

[a] Diagnosis and date of diagnosis are match
[b] Cardiovascular disease discharge record found in the LMR registry, but does not match completely with BHN intervention
[c] No cardiovascular disease discharge record found in the LMR registry

Linkage for group (1), who's partners' names in the NCR and BHN match, clearly show the best linkage outcome with the lowest percentage of surgical interventions not found in the DHD registry. Linkages for the groups (2), NCR's partner's name is filled, BHN's partner's name is empty, and (4), both partner's names are filled, but do not match, demonstrate to be much worse, with up to 73% of the surgical interventions not found in the DHD registry. Group (3), NCR's partner's name is empty, BHN's partner's name is filled, seems to be a mediocre group, not performing great, but not as bad as groups 2 and 4. This can be explained by the fact that in earlier years, NCR did not register partner's names. A missing partner's name in the NCR therefore does not seem to be equal to a missing partner's name in the BHN registry, where partner's names have always been registered. A missing partner's name in BHN probably

means that the patient was not married at the time, while a missing partner's name in the NCR can mean either no partner or simply not registered.

Data on other types of cancer

Survival of DCIS is high (96% at 5 years), this means that DCIS patients rarely die from breast cancer without a second, invasive breast cancer diagnosis. Although the analysis of breast cancer survival in DCIS patients was not our primary objective, these data provided us with an excellent opportunity to compare our data to what is known in the literature. It also proved to be an important check, as we saw more breast cancer deaths than expected, but also more deaths from other cancers that, according to our data, were not diagnosed in these patients. This indicated that there probably was an error in the linked dataset.

With an anonymised dataset it is difficult to identify the source of the problem, as we are not able to provide any of the involved parties with identifiable examples and we are the only ones with the combined data from all registries. Fortunately, we were able to allocate the problem in this case. There appeared to have been a misunderstanding with regard to the requested patient selection. We wished to receive all patients who had been diagnosed with DCIS or invasive breast cancer as their first primary neoplasm, i.e. patients without a history of a cancer prior to their DCIS or invasive breast cancer. Yet the NCR had understood we wanted all patients with a first DCIS or invasive breast cancer, irrespective of a prior history of cancer. Hence, we initially received data from DCIS patients (and invasive breast cancer patients) who had been diagnosed with other malignancies (including breast cancer in the case of DCIS) before their DCIS/invasive breast cancer. Consequently, patient selection and all subsequent linkages had to be performed again. This renders it clear how important it is to perform several data checks in an anonymised database even if cumbersome.

Discussion

Constructing a cohort with outcome data solely based on linkages is a complicated and time consuming process, but offers numerous possibilities of creating large and interesting cohorts. In our case, we were able to construct a unique, large, population-based cohort of breast cancer survivors and were able to take into account both cardiovascular morbidity and mortality by performing linkages with two population-based registries with CVD information and the national cause of death registry. From starting the process of acquiring the data to completion of all linkages took over two years, possibly because we were the first to undertake these linkages.

During the process of data collection we encountered several difficulties and problems, especially since this was the first time a cohort with multiple disease outcomes was constructed solely by linkage of the current data sources. First of all, starting up and making agreements for the data collection with the many different parties took quite some time. The protocol had to be reviewed by all boards and new contracts had to be established and checked by the other parties. Secondly, we were totally dependent on the other parties on performing the linkages, which meant trusting their abilities and adapting to their schedules.

In fear of the possibility of identification of the patients, none of the parties wanted the identification codes of the other parties to be sent on to the next party. Thus, each party removed the other parties'

identification code and added a new identification code to the data. As a consequence, no shortcuts were possible in the data collection path. Any additional data and edits had to go through the entire route.

During the linkage process, several different parties discovered that errors occurred in the identifying variables in their registry: patient names and partner's names had been mixed up or day and month of birth were switched. This was usually discovered during the linkage, when a certain hospital/institute had much fewer matches than the others.

Another problem that arose – which was discovered by chance – was that with all the different parties and hence the different computer programs used to read the data and perform linkages, data formats were sometimes changed, resulting in the loss of data. Because we did not have access to the original data, we did not know how many data cells should have been filled and with exactly what data. The parties had provided numbers of patients and number of matching record pairs, yet did not provide the number of filled cells per variable. We therefore did not notice that some third and fourth malignancies were missing, at first. This only became apparent when we noticed that one of the other variables contained incomplete words. It is therefore recommendable to request a detailed data description of all variables prior to linkage.

Unfortunately, 9.7% of our cohort could not be linked with the DHD due to the fact that these patients were not uniquely identifiable. The main disadvantage here is a loss of numbers. There is, however, no reason to assume difference in our outcome measure - CVD incidence - between identified and non-identified patients. We checked this by comparing patient characteristics and the frequency of cardiovascular surgical interventions between identified and non-identified patients, and no differences were found.

A second disadvantage of the rather limited set of linkage variables sex, date of birth and postal code, is that linkage is only possible for the periods in time during which a person is uniquely identifiable using these variables. This means that there may have been periods during follow-up for which no information on our disease outcome is available. We chose to only take into account the first period during which a person was uniquely identifiable, thereby decreasing the possibilities for such gaps. However, it is still possible that a person was admitted to the hospital for CVD without our knowledge if there was a period after breast cancer diagnosis during which that person was not uniquely identifiable.

Cardiovascular surgical interventions always include a hospital admission of at least one day. Therefore, in theory, all interventions provided by BHN should be present in the DHD database. Contrary to BHN, which is nearly complete since 2001, the completeness of the DHD has declined since 2003 from 99.4% to 87.3% in 2009. This decreased completeness is caused by a decline in participation in the DHD registry among hospitals. Yet, because hospital admissions for general CVDs are indiscriminately distributed among hospitals, the effect of the incompleteness for general CVDs is thought to be random. Only serious cardiovascular surgical interventions are limited to certain hospitals in the Netherlands. Such interventions are registered by BHN. Therefore, linkage with both population-based registries ensured a high coverage of CVD with only some random incompleteness of the less severe CVDs.

References

1. van der Sanden GA, Coebergh JW, Schouten LJ, et al. Cancer incidence in the Netherlands in 1989 and 1990: first results of the nation- wide Netherlands cancer registry. Coordinating Committee for Regional Cancer Registries. Eur J Cancer. 1995;31A(11):1822–1829.
2. F, Groenwold RHH, ter Burg WJ, et al. Nederlandse hartchirurgie over de periode 1995–2009 en de prognose tot 2020. In: Vaartjes I, van Dis I, Visseren FLJ, et al., eds. Hart- en vaatziekten in Nederland 2011, cijfers over leefstijl- en risicofactoren, ziekte en sterfte. Den Haag: Hartstichting; 2011:31–41.

Using record linkage to construct a population-based cancer patient cohort with multiple disease outcomes

PART II

Methods and experiences of the Netherlands Cancer Institute in the construction of an enriched population-based cohort of women diagnosed with ductal carcinoma in-situ of the breast in the Netherlands, linked with the nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA) and the Netherlands Breast Screening Database (IBOB).

Introduction

For many studies it is of interest to enrich a cohort of cancer patients derived from the cancer registry with pathological characteristics of the cancers, or with information on subsequent diseases/surgeries occurring during follow-up, as registered by the Dutch nationwide network and registry of histo- and cytopathology (PALGA). Furthermore, linkage with PALGA provides the unique opportunity to obtain access to tumour tissue blocks in the individual pathology laboratories, enabling assessment of tumour markers or more advanced molecular genetic profiling.

For a study aiming to find a gene expression profile predicting risk of subsequent ipsilateral invasive breast cancer in patients with ductal carcinoma in-situ (DCIS) of the breast, we constructed a large cohort of women diagnosed with pure DCIS in the Netherlands with detailed follow-up information, data regarding breast screening history and causes of death. This cohort was subsequently used to select relevant subgroups of women for whom archived tumour tissue was to be obtained through PALGA. The initial data collection for this cohort was performed solely by linkages of different population-based registries: a time-consuming process that involved many different organizations and regulatory boards safeguarding the privacy of the patients involved.

In this report we discuss the linkage procedures and our experiences with the process of constructing a cohort by linkages only. We also address the retrieval of selected archived tumour material through PALGA.

Box 2. The different parties involved in the construction of the second cohort.

The Netherlands Cancer Registry *(NCR; Integraal Kankercentrum Nederland, de Nederlandse Kankerregistratie: http://cijfersoverkanker.nl/gegevens-aanvragen-16.html):* The NCR is a population-based nationwide cancer registry and has a coverage of at least 96% of invasive malignant neoplasms and selected non-invasive cancers occurring in the Netherlands since 1989 [8]. The NCR receives regular notification of new cancer diagnosis from PALGA; yearly linkage of the NCR with the Dutch hospital discharge diagnosis registry and more recently with hospital-based DBC (diagnosis-treatment combinations) lists complete case-ascertainment.

The nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA, http://www.palga.nl/gegevensaanvragen/aanvraag.html): The PALGA database contains abstracts of all Dutch pathology reports; consisting of encrypted patient identification, a summary of the pathology report, diagnostic terms in line with SNOMED terminology. Currently about 64 million abstracts of more than 12 million patients are stored and each year more than 2 million abstracts are added. PALGA has complete coverage of all Dutch pathology laboratories since 1990. PALGA does contain pathology reports from before 1990, from pathology laboratories already participating in PALGA before 1990, but these data are incomplete. Through PALGA it is possible to retrieve tumor tissue blocks from the pathology laboratories to perform additional analyses, e.g., determining immunohistochemical markers, molecular genetic tests.

The Netherlands Breast Screening program (IBOB, Informatiesysteem Bevolkingsonderzoek Borstkanker)*:* The IBOB database contains person data and limited medical data for women who participate(d) in the nationwide Dutch breast screening program (which was started in 1990 and reached complete coverage of the target population in 1997). IBOB preserves these records only for a period of 15 years, counted from the year in which the screened woman, based on her attained age, no longer belongs to the target population for the breast screening program. On request, screening records may be destroyed earlier.

Statistics Netherlands (CBS, Centraal Bureau voor de Statistiek: http://www.cbs.nl/nl-NL/menu/informatie/beleid/zelf-onderzoeken/default.htm): Statistics Netherlands keeps record of many different statistics (so called microdata), including the nation-wide cause of death registry. Individual researchers can get access to these microdata, either at the bureau of Statistics Netherlands or via remote access to Statistics Netherlands microdata, under strict conditions, when the director-general of statistics has

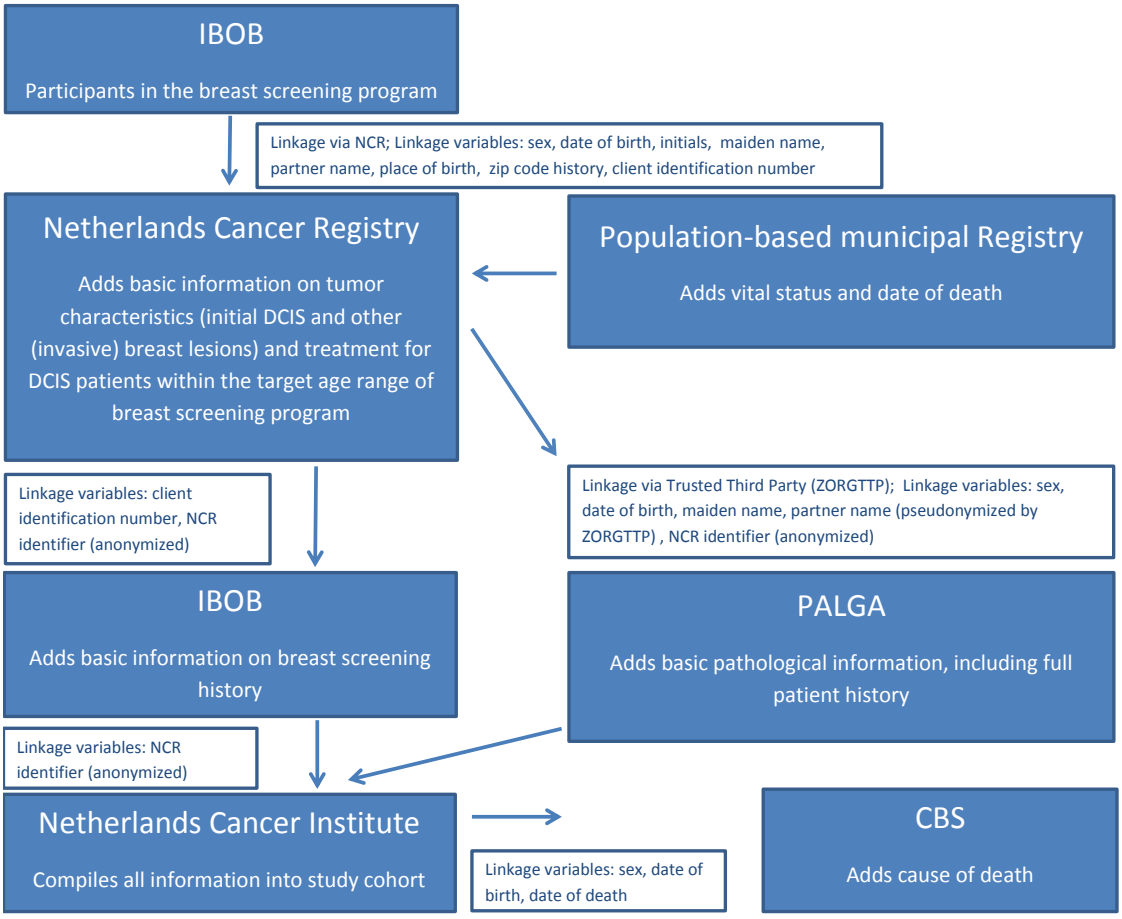
authorized the institution to which the researcher is affiliated to use the microdata facilities at Statistics Netherlands.

Methods

Data collection

The routing of the data collection and linkages is shown schematically in Figure 2. Our project and data/linkage requests were reviewed and approved by the internal review board of the NCR (Commissie van Toezicht op de Nederlandse Kanker Registratie) the internal review board (Wetenschappelijke Raad) and the privacy committee of PALGA and the Board of the Dutch population screening initiatives (Bestuur van de Facilitaire Samenwerking Bevolkingsonderzoeken (FSB)).

Figure 2. Data collection route for establishing a population-based breast cancer cohort of women diagnosed with pure DCIS as first neoplasm.



From the IBOB database, identifiers (sex, date of birth, initials, birth name, partner's name, place of birth, date of death, postal code history, IBOB client number) of all women who participated in the breast screening for at least one screening round were selected for linkage with the NCR and sent to the NCR. The database sent to the NCR was in long format, with one record for each postal code. The NCR

subsequently selected a cohort comprising all women diagnosed with ductal carcinoma in situ of the breast (DCIS) as their first neoplasm between 1989 and 2005 in the Netherlands. In total, this cohort consisted of 12,309 patients. The NCR acquires vital status and date of death through an annual linkage with the population-based municipal personal records database. This cohort was linked to the data provided from the IBOB database. For all women within the cohort who could be linked to the IBOB data, the IBOB client number and a unique NCR reference number were returned to IBOB. IBOB subsequently added the screening history of these women to the NCR reference number. For about 10% of the women who had participated in the breast screening program no clinical data could be added due to the fact that the screening records for these women were not preserved (it could not be determined whether these women had participated in the breast screening program). These data were successively delivered to the researchers at the Netherlands Cancer Institute

Simultaneously, the NCR provided the DCIS cohort for linkage with the PALGA database to ZorgTTP, which acts as a trusted third party for all NCR-PALGA linkages (https://www.zorgttp.nl). The NCR-cohort comprised clinical information regarding the DCIS lesion and all subsequent (breast) neoplasms following after the index DCIS. The linkage variables comprised sex, date of birth, birth name (at least first four letters) and partner's last name, as well as the unique NCR reference number. ZorgTTP pseudonymised (encrypted) the personal identifiers in this cohort and forwarded this encrypted cohort to PALGA for further linkage to the PALGA database. PALGA subsequently added all pathology abstracts from three months prior to the DCIS diagnosis to the database (multiple records per NCR reference number). This enriched NCR cohort, without personal identifiers, was successively delivered to the researchers at the Netherlands Cancer Institute.

At the Netherlands Cancer Institute the breast screening history was linked to the enriched NCR cohort. This cohort was uploaded to Statistics Netherlands for linkage with the cause of death registry Unlike the procedure described in the first part of this paper, we did not use the RIN- number as a linkage variable, but had the linkage executed using the near 100% unique combination of date of birth, date of death and sex (a so called "maatwerk-koppeling", or tailored linkage). After this linkage, Statistics Netherlands removed all identifying variables (including day of birth), encrypted the NCR reference number and made the anonymised data available to researchers at the Netherlands Cancer Institute via the remote access connection on a dedicated computer at the Netherlands Cancer Institute.

Performed data checks

The NCR and IBOB provided the data in a wide data structure, with only one record a person. PALGA, however, has a long data structures with multiple records per person, i.e. one record for every pathology report submitted for that person. The PALGA data was scrutinized to ensure that de initial lesion reported by the NCR was indeed pure DCIS and no invasive component had been missed by the NCR. Furthermore, treatment details in the NCR were checked against PALGA reports to determine the type of surgery: we verified based on PALGA data whether patients truly had only been treated with wide local excision and not subsequently received mastectomy and we checked PALGA to determine type of surgery for patients for whom the type of surgery was undetermined in the NCR data). Additionally, PALGA data were used to determine whether DCIS or invasive recurrences had been missed by the NCR and to determine whether patients may have undergone prophylactic mastectomies.

Retrieval of archived tumour tissue through PALGA

Selecting patients additional tumour material

To study risk of developing a subsequent invasive breast cancer in the same breast where the DCIS was located, we set up a case-control study. From the NCR-PALGA cohort we selected all patients who were solely treated with a wide local excision (no additional radiotherapy or systemic treatment) who subsequently developed an invasive ipsilateral breast cancer (316 patients: cases). For each of these patients 4 women were selected, matched to the case on age, who were also treated with wide local excision only but who had not developed an invasive ipsilateral breast cancer in the interval, calculated from DCIS diagnosis, in which the case developed her ipsilateral invasive breast cancer (1264 patients: controls). For all selected controls we checked whether PALGA contained a DCIS excerpt (coded summary of the pathology report which includes the conclusion of the pathology report in free text) and whether the patient had not undergone a mastectomy (prophylactic or for DCIS recurrence) in the interval in which the case developed her invasive breast cancer. Subsequently we selected for all cases and controls the PALGA excerpts for the DCIS resection; only in case the excerpt of the resection mentioned no tumour or when no excerpt for a resection was available, the excerpt of the previous DCIS biopsy was selected. For the cases, we also selected the excerpt detailing the pathological examination of the resected subsequent ipsilateral invasive breast cancer. For each case we have requested archived material of at least two controls.

A file containing for each patient and for each of the selected excerpts a PALGA administrative number, the PALGA excerpt number and the date on which the specified material was received by PALGA was subsequently sent to PALGA. This file contained a total of 1371 DCIS excerpts of 1101 patients and 336 invasive breast cancer excerpts of 320 patients; for 4 patients PALGA showed an invasive ipsilateral breast cancer not (yet) included in the NCR). Additionally, we uploaded a study proposal and a request for the pathology laboratories where the material was located, to send us the original pathology reports, the original tissue sections (coupes) and the tumour blocks concerning the selected PALGA excerpts.

Request for tumour material by PALGA

The file with PALGA administrative number, the PALGA excerpt number and the date of receiving the specified material by PALGA was linked to the PALGA database and subsequently the PALGA T-numbers (tissue block numbers) and the pathology laboratory where the material was processed were added to the file.

Subsequently, PALGA sent out our uploaded data request and study proposal as well as a pathology laboratory specific list with DCIS T-numbers and a list with T-numbers for invasive tumours to all Dutch pathology laboratories where one or more of the selected tissue blocks were located. We received a file with a list of the pathology labs that were approached by PALGA and the number of pathology reports and tissue blocks requested in each of the pathology laboratories (we did not receive T-numbers).

Response of pathology laboratories

Basically follow-up actions remain limited to waiting for the material. We mailed or phoned the labs to remind them of our request. A few times lists were lost and we requested PALGA to resend the lists to pathology laboratories. The material we received was checked for completeness (pathology reports, the original tissue sections (coupes) and the tumor blocks) and archived. As the quality of the original tissue sections proved to be very poor for our purpose (revision of the original pathology), we decided to stop requesting the original tissue coupes.

Up to December 5th, 2014, we had received the requested materials from 45 of the 58 selected pathology laboratories (78%). Two other labs agreed to co-operate but had not yet sent in the requested materials. Three pathology laboratories refused participation (about 1% of all excerpts) and samples from eight labs were not received for unspecified reasons (figure 3).

We received material for 82% (1127) of the selected DCIS (3% pending) excerpts as well as material for 82% (274) of the selected invasive breast cancers (4% pending, figure 4).

Figure 3. Response of pathology laboratories to the request for tissue sections.

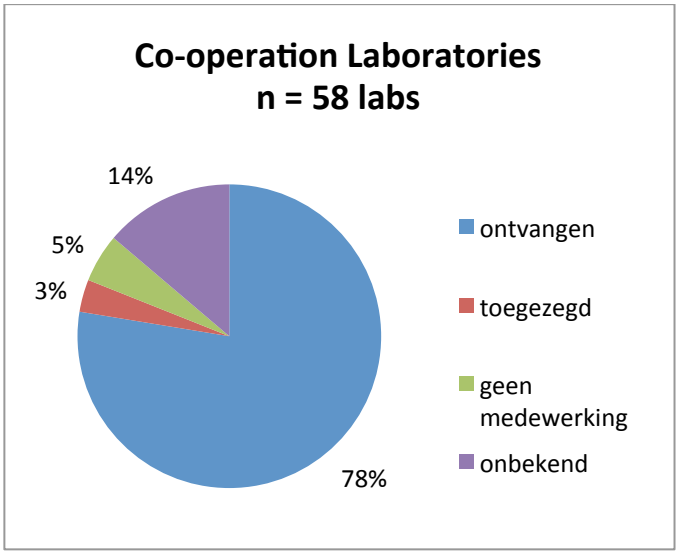
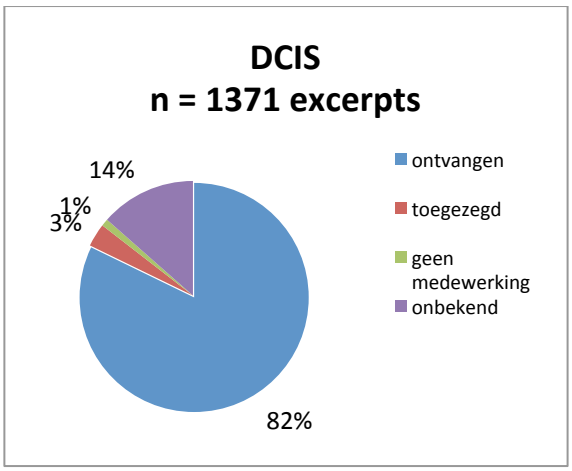
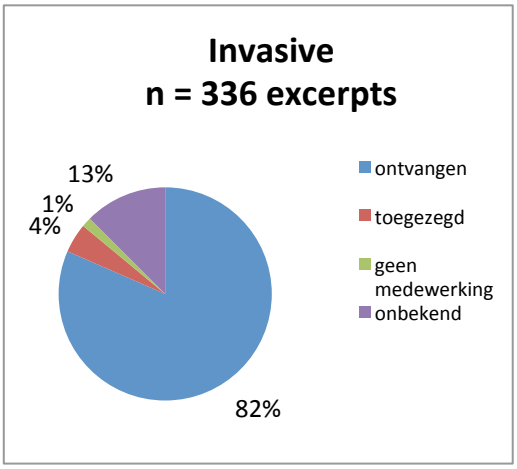


Figure 4. The retrieved DCIS and invasive breast cancer excerpts as a percentage of the requested number of excerpts.





Discussion

Enriching this nationwide population-based DCIS cohort by linking various data sources provided a wealth of information. The NCR-PALGA linkage additionally provided opportunities for collecting

complete pathology reports and archived tumour material through the Dutch pathology laboratories. PALGA's support is instrumental in this undertaking.

Constructing this cohort solely by linkages was a time consuming process, however. From starting the process of acquiring the data to completion of all linkages took over two years. First of all, our research protocol had to be reviewed by four review boards and several contracts had to be prepared and signed between the involved parties. As we stated in the first part of this report, we largely depended on the various parties involved for performing the linkages, which meant trusting their abilities and adapting to their schedules. The IBOB-NCR linkage proved to be especially time-consuming due to the fact that some issues arose with regard to the period for which the IBOB was allowed to preserve information on the women they had invited for screening in the past. With regard to the linkage with our DCIS cohort, internal analysis by IBOB showed that for about 10% of the DCIS population, for which there was a positive match between the NCR database and the IBOB database on personal identifiers, all clinical data were already deleted. For these patients IBOB could not provide data to external parties.

To exclude the possibility of identification of any of the patients involved, we as researchers were not allowed to have any personal identifiers in the final database, allowing very limited room for checking the data.

We are currently finalising the process of collecting archived tumour material for a subset of all DCIS patients for further immunohistochemical and molecular analysis. Although the process was a lengthy one, the possibility of linking basic cancer patient information from the nationwide population-based NCR with data from the nationwide PALGA registry, and especially the opportunity to collect tumour tissues through PALGA, provides a unique environment for clinical and molecular epidemiological research, which is rivalled by only few countries worldwide.