Identification of patients with incident cancers using administrative registry data

Mette Bach Larsen[1, 2], Henry Jensen[1, 2], Rikke Pilegaard Hansen[1, 2], Frede Olesen[1] & Peter Vedsted[1, 2]

ABSTRACT

INTRODUCTION: On-time identification of incident cancer patients is important in cancer research to ensure quality in cancer treatment and care. Nevertheless, the Danish Cancer Registry (DCR) is updated on an annual basis rather than continuously, and no standardised algorithm exists to enable sampling from administrative data which are updated on a monthly basis. The aim of this study was to develop and validate an algorithm for on-time sampling of incident cancer patients based on administrative data.

MATERIAL AND METHODS: The study was based on registry and questionnaire data from incident cancer patients' general practitioners (GPs). An algorithm for on-time sampling of incident cancer patients was developed and validated in 2008 (12,747 patients) and further developed and validated in 2010 (7,996 patients). Questionnaire data from the GPs and data from the DCR were used as golden standards. The completeness over time of the 2010 cohort was evaluated.

RESULTS: Further development of the 2008 algorithm into the 2010 algorithm increased its positive predictive value (PPV) to 95.0%. The PPV of a patient from the 2010 cohort being registered in the DCR was 97.4%. The 2010 algorithm displayed a completeness of 60% in the first month and 95% after four months.

CONCLUSION: A valid and cost-saving algorithm for on-time sampling of incident cancer patients has been developed with great potential for research and quality assurance.

FUNDING: This work was funded by the Danish Cancer Society and the Novo Nordisk Foundation.

TRIAL REGISTRATION: not relevant.

Cancer is a major health-care burden in Denmark with a life-time risk of 33%, approximately 35,500 new cases per year and 15,500 annual deaths [1, 2]. Many resources are consequently allocated to research and quality improvement of the entire cancer care pathway from early symptoms, diagnosis and treatment through rehabilitation or palliation. For many of these purposes, collection of data at the time of diagnosis (or rapidly thereafter) is crucial. Furthermore, initiation of interventions at the onset of the cancer pathway is also often needed. Collection of such data is frequently a laborious and time-consuming activity.

A number of registries collect data on cancer patients by using the Danish civil registration number to identify each patient. Denmark hosts the world's oldest cancer registry, the Danish Cancer Registry (DCR), which contains data on the incidence of cancer throughout Denmark since 1943. However, on-time data cannot be extracted from the DCR since the registry is updated only on an annual basis [3]. Therefore, on-time identification of incident cancer patients must be based on administrative registries, but no standardised algorithm has been developed to allow this.

The aim of the present paper was to develop and validate a registry-based algorithm for on-time sampling of incident cancer patients and to describe the completeness of a cohort of incident cancer patients identified by the algorithm.

MATERIAL AND METHODS

Setting

The study was carried out among general practitioners (GPs) in the Region of Central Jutland and the Region of Southern Denmark. More than 98% of Danish citizens are registered with a GP, and the diagnostic pathway is initiated in general practice for approximately 85% of all cancer patients [4].

Data sources

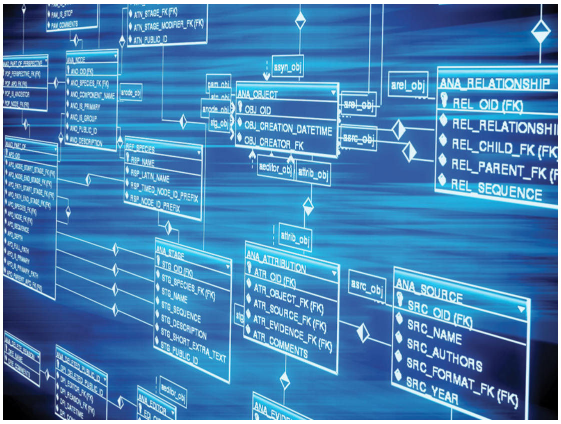
The purpose of using the regional Patient Administrative Systems (PASs) was to identify patients in the 2008 cohort. PAS is to collect administrative information on hospital activities. PAS comprises information on every patient contact with the hospitals including patients' civil registration number, dates of admission and discharge and diagnosis classified according to the International Classification of Diseases (ICD 10). Patient registration is made in accordance with national guidelines [5] assuring minimum regional differences. To provide data for the National Patient Registry (NPR), the hospitals are committed to update PAS for the previous month by the 10th of each month and to report these data to the NPR [6].

The NPR is a national database run by the Danish Health and Medicines Authority. The registry comprises information from the regional PASs. The NPR was originally developed to monitor hospital activities, but has also served as a basis for payment of the hospitals since 2000. The NPR is also used for medical research such as

ORIGINAL ARTICLE

1) Research Unit for General Practice and Department of Public Health, Aarhus University
2) Research Centre for Cancer Diagnosis in Primary Care, Aarhus University

Dan Med J 2014;61(2):A4777



Using administrative data in a research algorithm.

epidemiological studies [6, 7]. Any registration of cancer in the NPR triggers a duty of notification to the DCR. Thus, registration of a new cancer diagnosis requires the additional code AZCA1 which indicates that this diagnosis is registered for the first time [5].

The DCR is a national research registry designed to collect and process data on incident cancer cases. The DCR contains information on e.g. date of diagnosis, tumour topography, morphology and spreading. Due to comprehensive quality control, the DCR is updated on an annual basis. Within a year, almost 90% of the tumours in the DCR are validated [6, 8, 9].

The initial GP questionnaire was developed in 2007 based on literature and research group experience from prior studies into patients' diagnostic pathways [10, 11]. Within one month after the patient was discharged from the hospital with a cancer diagnosis, the GPs received a questionnaire requesting information on whether the patient had cancer, if the cancer had been diagnosed within the previous six months (including date of diagnosis) and whether the registry-based diagnosis was correct. For the 2008 cohort, non-responding GPs received a reminder after three weeks. For the 2010 cohort, a similar questionnaire was sent to GPs with a reminder after six weeks.

Definition of an incident cancer patient

An incident cancer patient was defined by the following characteristics: 1) discharged from a hospital with cancer, 2) no prior history of cancer, 3) only one cancer diagnosis present and 4) the cancer was diagnosed within six months of inclusion. Patients with non-melanoma skin cancers (C44) were excluded as were patients younger than 18 years of age.

Developing the sampling algorithms

The first sampling algorithm was developed in the spring of 2007 based on expert meetings with persons holding administrative responsibility for registering cancer patients, persons in charge of handling output from registers and people with considerable research experience. Patients were sampled from PASs in the Region of Central Jutland and the Region of Southern Denmark based on discharge date, diagnosis and the additional code AZCA1 [5]. Patients were sampled on the 15th of each month, and data on all patients registered during the preceding month were collected. Patients were not considered incident if they were already registered on a national list of all cancer diagnoses from 1994 onwards. The monthly sampling continued for a one-year period from 1 October 2007.

The algorithm proved to be incomplete for two reasons. First, some patients were registered later than one month after their diagnosis and were missed because the algorithm only sampled one month back. Second, the AZCA1 code was not used consistently, which implies that not all eligible patients were included. An additional sampling was therefore performed 11 months after the sampling period. This sampling procedure did not state the AZCA1 code as an inclusion criterion, and all patients were sampled simultaneously for the entire study period. The 2008 cohort consists of samples one and two combined (**Figure 1**).

Based on the experiences from the 2008 cohort, an improved algorithm was developed for sampling of the 2010 cohort. The main differences between the two algorithms were that monthly updates included patients from the previous months to ensure incorporation of patients who had been registered late and that a prior history of cancer was based on the DCR until 31 December 2008 and on the NPR for 2009.

Finally, diagnoses classified as D37-D48 were excluded in this cohort (approximately 3% in the 2008 cohort). Patients were sampled from 1 January 2010 to 31 October 2011 (Figure 1).

Validating the sampling algorithms

In order to use a non-registry based golden standard, the sampling algorithms were validated using information on diagnosis and date of diagnosis obtained from the patients' GPs. Because of the high validity of the DCR, we also used the DCR as a golden standard. This was possible one year after patient inclusion for both cohorts.

Completeness of the 2010 cohort

We evaluated the completeness of the 2010 cohort over time, stating how many months it would require to have a complete cohort of incident cancer patients from July 2010. Completeness was defined as the time when all sampled cancer patients were registered in the NPR. The level of completeness was measured as the cumulated monthly proportion of cancer patients sampled.

Analysis

Positive predictive values (PPVs) for sampling an incident cancer patient were calculated using the GP and the DCR as golden standards. Further, differences between patients in the cohorts and patients in the DCR were tested to state whether the cohorts were a representative sample of the DCR.

The completeness of the sampling of the 2010 cohort was tested by comparing the number of cancer patients sampled each month over time with the overall number of cancer patients sampled.

Statistical significance was defined as a probability of 5% or less. Analyses were made using Stata 11.2.

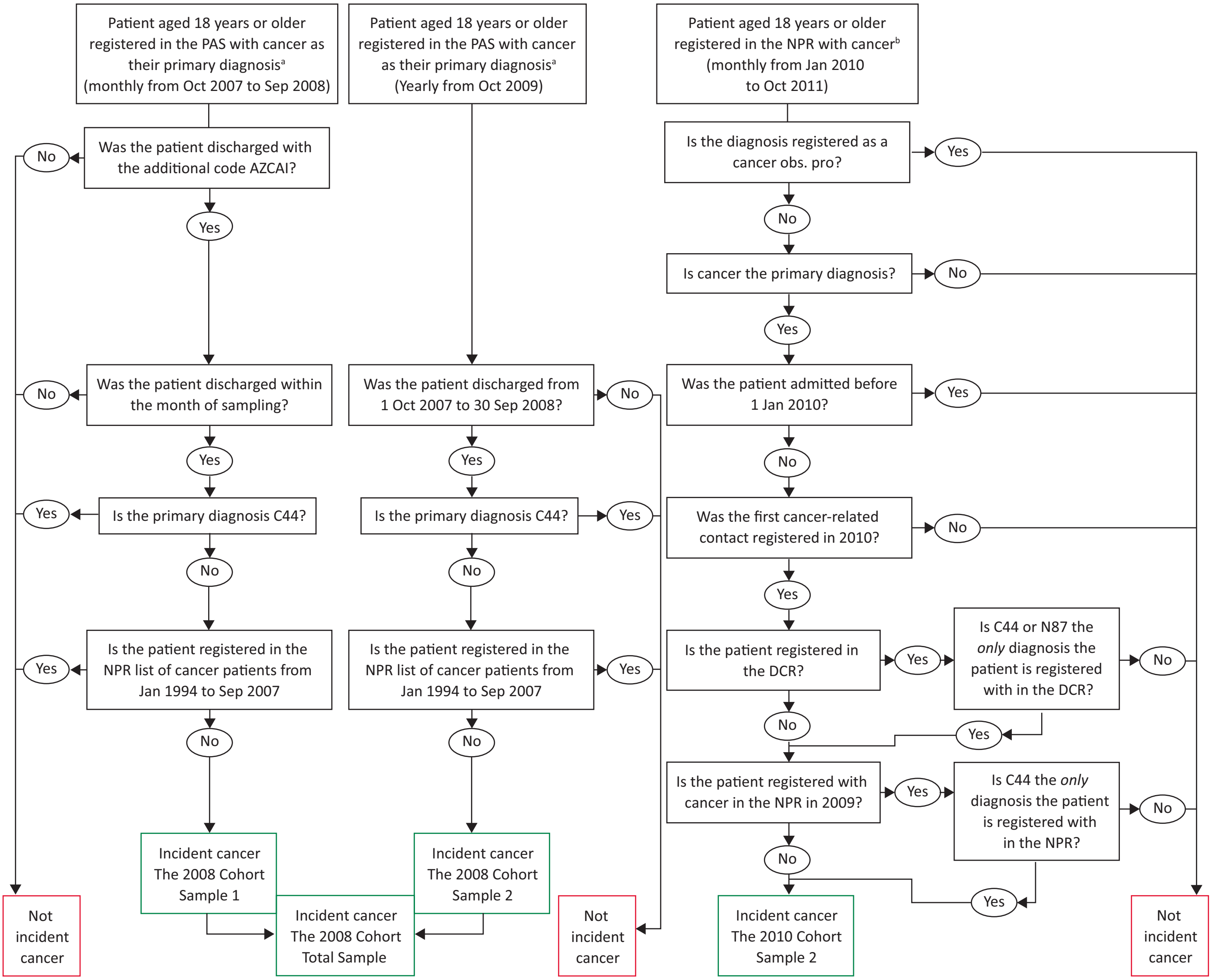
Ethics approval

According to the Research Ethics Committee of the Region of Central Jutland, the Danish acts on research ethics review of health research project do not apply to this project. The study was approved by the Danish Data Protection Agency and the Danish Health and Medicines Authority.

Trial registration: not relevant.

FIGURE 1

The sampling algorithms.


Patient aged 18 years or older registered in the PAS with cancer as their primary diagnosis[a] (monthly from Oct 2007 to Sep 2008)
Patient aged 18 years or older registered in the PAS with cancer as their primary diagnosis[a] (Yearly from Oct 2009)
Patient aged 18 years or older registered in the NPR with cancer[b] (monthly from Jan 2010 to Oct 2011)
Was the patient discharged with the additional code AZCAI?
No
Yes
Was the patient discharged within the month of sampling?
No
Yes
Is the primary diagnosis C44?
Yes
No
Is the patient registered in the NPR list of cancer patients from Jan 1994 to Sep 2007
Yes
No
Incident cancer The 2008 Cohort Sample 1
Was the patient discharged from 1 Oct 2007 to 30 Sep 2008?
No
Yes
Is the primary diagnosis C44?
Yes
No
Is the patient registered in the NPR list of cancer patients from Jan 1994 to Sep 2007
Yes
No
Incident cancer The 2008 Cohort Sample 2
Incident cancer The 2008 Cohort Total Sample
Not incident cancer
Is the diagnosis registered as a cancer obs. pro?
Yes
No
Is cancer the primary diagnosis?
No
Yes
Was the patient admitted before 1 Jan 2010?
Yes
No
Was the first cancer-related contact registered in 2010?
No
Yes
Is the patient registered in the DCR?
Yes
Is C44 or N87 the only diagnosis the patient is registered with in the DCR?
No
Yes
No
Is the patient registered with cancer in the NPR in 2009?
Yes
Is C44 the only diagnosis the patient is registered with in the NPR?
No
Yes
No
Incident cancer The 2010 Cohort Sample 2
Not incident cancer
Not incident cancer

DCR = Danish Cancer Registry; NPR = Danish National Patient Registry; PAS = Patient Administrative System.
a) Cancer diagnoses C00.0-C99.9 + D37-D48 according to the ICD 10. b) Cancer diagnoses C00.0-C99.9 according to the ICD 10.

RESULTS

In the 2008 cohort, a total of 10,262 out of 12,747 GP questionnaires were completed (80.5%). In the 2010 cohort, a total of 5,711 out of 7,996 GPs filled in the questionnaire (71.4%). For both cohorts, patients with a non-responding GP were more likely to be older men diagnosed with prostate cancer ($p < 0.05$).

Validating the cohorts

Further development of the 2008 algorithm into the 2010 algorithm increased the PPV of sampling an incident cancer from 83.5% to 95.0% (**Table 1**). The PPV of a patient from the 2010 cohort being DCR-registered was 97.4%. Of the 211 patients who were not registered in the DCR, the GP verified that the patient had cancer in

TABLE 1

Positive predictive values in percentage of sampling a cancer patient and an incident cacer patient.

	The 2008 cohort					
	sample 1		total sample		The 2010 cohort (N = 5,556)	
	n	PPV (95% CI)	n	PPV (95% CI)	n	PPV (95% CI)
Cancer, all[a]	6,587	97.9 (97.3-98.0)	9,928	96.7 (96.4-97.1)	5,491	98.8 (98.5-99.1)
Included cancer[b]						
Lung cancer	781	99.5 (98.7-99.9)	1,214	99.1 (98.4-99.6)	677	98.7 (97.5-99.4)
Colorectal cancer	881	98.9 (97.9-99.5)	1,279	98.7 (97.9-99.2)	737	99.1 (98.1-99.6)
Prostate cancer	830	99.3 (98.4-99.7)	1,302	98.9 (98.2-99.4)	745	99.2 (98.3-99.7)
Malignant melanoma	331	97.4 (95.0-98.8)	467	97.3 (95.4-98.6)	275	97.5 (94.9-99.0)
Breast	1,305	99.6 (99.1-99.9)	1,646	99.6 (99.1-99.8)	954	99.8 (99.2-99.9)
Other	2,458	95.1 (94.2-95.9)	4,008	93.4 (92.6-94.1)	2,075	97.1 (96.3-97.8)
Total	6,586	97.6 (97.2-98.0)	9,916	96.6 (96.3-97.0)	5,463	98.3 (97.9-98.6)
Incident cancer[c]						
Lung cancer	732	93.2 (91.3-94.9)	1,125	91.8 (90.2-93.3)	729	95.4 (93.6-96.9)
Colorectal cancer	805	90.3 (88.2-92.2)	1,159	89.4 (87.6-91.1)	671	94.3 (92.4-95.8)
Prostate cancer	609	72.8 (70.0-75.8)	905	68.8 (66.2-71.3)	644	83.4 (80.6-86.0)
Malignant melanoma	313	92.1 (88.7-94.7)	435	90.6 (87.7-93.1)	267	93.0 (89.4-95.7)
Breast	1,182	90.2 (88.5-91.8)	1,466	88.7 (87.1-90.2)	946	96.4 (95.1-97.5)
Other	2,175	84.2 (82.7-85.6)	3,453	80.5 (79.2-81.6)	2,020	92.0 (90.8-93.1)
Total	5,816	86.3 (85.4-87.1)	8,543	83.5 (82.5-84.0)	5,277	95.0 (94.4-95.5)

CI = confidence interval; GP = general practitioner; PPV = positive predictive value.
a) The GP confirms that the patient has cancer.
b) The GP confirms that the patient has a cancer included in the study.
c) The GP confirms that the patient has an incident cancer as defined in the study.

TABLE 2

Validating the cohorts based on the Danish Cancer Registry.

	Total 2008			2010		
	cohort, n (%) (N = 12,747)	DCR, n (%) (N = 10,948)	difference: cohort ⍰ DCR, %-points (95% CI)	cohort, n (%) (N = 7,996)	DCR, n (%) (N = 26,659)	difference: cohort ⍰ DCR, %-points (95% CI)
Sex						
Male	6,394 (50.2)	5,514 (50.4)	–0.2 (–2.0-1.6)	4,160 (52.0)	13,603 (51.2)	0.8 (–0.4-2.1)
Female	6,353 (49.8)	5,434 (49.6)	0.2 (–1.6-2.0)	3,836 (48.0)	12,966 (48.8)	–0.8 (–2.1-0.4)
Age						
18-49 yrs	1,396 (11.0)	1,263 (11.4)	–0.4 (–2.8-2.0)	831 (10.4)	3,100 (11.7)	–1.3 (–2.0-0.5)
50-69 yrs	5,760 (45.2)	5,181 (47.3)	–2.1 (–4.0- –0.2)	3,771 (47.2)	12,486 (46.8)	0.2 (–1.1-1.4)
≥ 70 yrs	5,591 (43.9)	4,504 (41.1)	2.8 (0.9-4.7)	3,394 (42.5)	11,073 (41.5)	0.8 (–0.5-2.0)
Diagnosis						
Breast cancer	1,978 (15.2)	1,680 (15.2)	0.0 (–2.3-2.3)	1,314 (16.4)	4,396 (16.5)	–0.1 (–1.0-0.8)
Lung cancer	1,493 (11.7)	1,452 (13.3)	–1.6 (–4.0-0.8)	989 (12.4)	3,397 (12.8)	–0.4 (–1.2-0.4)
Colorectal cancer	1,603 (12.6)	1,440 (13.2)	–0.6 (-3.0-1.8)	1,054 (13.2)	3,489 (13.1)	0.0 (–0.8-0.9)
Prostate cancer	1,705 (13.4)	1,370 (12.5)	0.9 (–1.5-3.3)	1,143 (14.3)	3,824 (14.3)	–0.1 (–1.0-0.8)
Malignant melanoma	610 (4.8)	541 (4.9)	–0.1 (–2.6-2.4)	404 (5.1)	1,590 (6.0)	–0.9 (–1.5-0.4)
Other	5,358 (42.0)	4,465 (40.8)	1.2 (–0.8-3.2)	3,092 (38.7)	9,963 (37.4)	1.2 (–0.0-2.4)

CI = confidence interval; DCR = Danish Cancer Registry.

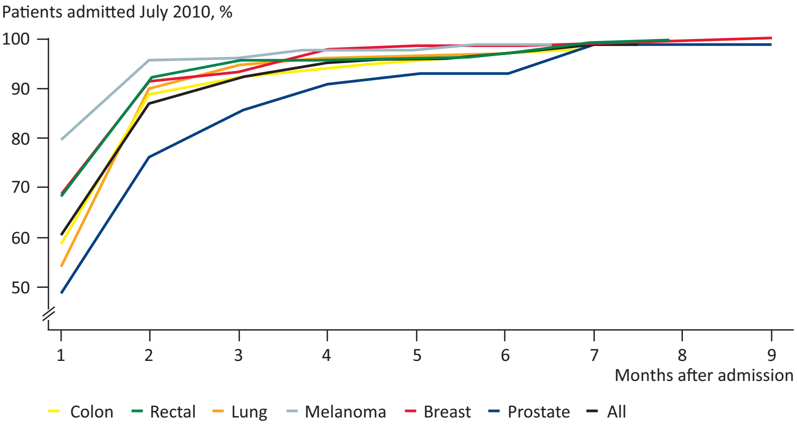
200 cases (94.8%) (2010 cohort). No statistically significant gender and diagnosis differences were observed between patients included in the study and patients registered in the DCR, whereas patients in the 2010 cohort were less likely to be young **(Table 2**).

Completeness over time

Overall, the completeness of registration of incident patients from June 2010 was 60.0% within the first month with variations between diagnoses; completeness was lowest for prostate cancer (49.0%) and highest for malignant melanoma (-79.5%). After four months, the overall completeness exceeded 95%, with variations from 90.5% (prostate cancer) to 98.1% (breast cancer). A minimum completeness of 95% was achieved within two to five months after admission, except for prostate cancer (**Figure 2**).

FIGURE 2

Completeness of included patients over time. Note that the graph is constrained to nine months, but the analysis is undertaken for all 14 months.



DISCUSSION

Main findings

The 2010 algorithm for on-time sampling of incident cancer patients was developed and tested. The algorithm reached a PPV of 95.0%. The PPV of a 2010 cohort patient being registered in the DCR was 97.4%. When the GP was used as the golden standard, 85.4% of the patients who were not registered in the DCR had a cancer diagnosis. Finally, the 2010 algorithm displayed a completeness of 60% in the first month and a completeness of 95% after four months.

Strengths and weaknesses

The greatest strength of this study is that the development of the sampling algorithm was documented and tested in a large-scale study. Furthermore, the considerable sample size improves the statistical precision of our findings.

The risk of selection bias was minimised by the registry-based sampling, provided that registered data were valid. A study from 1993 found that only 73% of the patients were registered with the correct diagnosis in the NPR [12]. Yet, consistent with our results, several recent studies conclude that minor misclassifications do exist in the NPR, especially where the coding practice is unclear. However, the misclassifications are non-systematic and do not influence the overall validity of the NPR data [6, 7, 13-17].

Information bias may be a risk due to GP recall bias. However, the GPs were specifically asked to base their answers on their medical records and discharge letters from the hospitals in order to minimise this possible bias. Furthermore, assuming that cancer cases are randomly distributed in the population, a GP will only see 8-10 new cancer patients a year [18]. This was also an important consideration when choosing the GP as a golden standard. The GPs are expected to have detailed knowledge on the few patients with cancer in their practice because of the severity of the disease and because they often initiate the diagnostic process. Therefore, the GPs were considered a suitable alternative to the registries. Even so, a risk of information bias does prevail, because we assume that the GPs will be more likely to respond if they know nothing about a patient's given cancer. Thus, the PPVs presented should be considered minimum estimates.

We found that some patients were not registered in the DCR, but were verified by their GP as having cancer. This could be explained by the DCR validation procedures, which determine that patients with discrepancies between registration details and pathology report are put on hold. The validation only showed minor differences in the distribution of age and cancer types; this indicates that no systematic differences were identified between patients included in the study and those registered in the DCR.

The use of completeness as a measure to estimate whether the sampling algorithm incorporated all incident cancer patients could be questioned. Completeness is used to estimate whether a database can be used to recruit the eligible population [19] and it is a measure that must be taken into account; it has been argued that completeness should reach 90% to ensure that a sample is representative of the studied population [20]. On this basis, the 2010 cohort can be considered a representative sample of all incident cancer patients.

The algorithm displayed relatively low PPVs for the sampling of incident prostate cancer patients. This discrepancy could be rooted in the long diagnostic pathway experienced by most prostate cancer patients, which may lead to delayed registration or misclassification. Further, the GPs are more likely to be non-responders when the patients have prostate cancer; this may imply that the questionnaire is not suitable for capturing the care pathway of these patients. Breast cancer displays the highest PPV which indicates that its diagnosis and registration is more straight-forward than for some of the other cancers. Furthermore, malignant melanoma shows a high degree of completeness already within a few months which indicates that registration of this diagnosis is done almost on-time.

Generalisability

Overall, the results of this study could easily be transferred to other studies of incident cancer patients. However, the definition of an incident cancer patient must be applicable to these studies. The high PPV in the 2010 sampling algorithm was partly achieved by excluding patients classified with the diagnoses D37-D48 (neoplasm of uncertain or unknown behaviour), and this may not be suitable for all projects. Yet, since only a minor part of the included cancer patients are affected, this is unlikely to influence the study results.

CONCLUSION

As stated by others, using administrative data in research holds a great potential along with many potential difficulties. Knowledge of the administrative data is likely to magnify the advantages and minimise the potential problems [20]. Now a valid and cost-saving algorithm for on-time sampling of incident cancer patients has been developed with much potential for future research and quality assurance.

Using the algorithm is a trade-off between on-time identification and high validity of diagnosis. Sampling within one month results in a completeness of 60%, whereas sampling over four months yields a completeness of 95%.

CORRESPONDENCE: *Mette Bach Larsen,* Forskningsenheden for Almen Praksis, Aarhus Universitet, Bartholins Allé, Bygn. 1260, 8000 Aarhus C, Denmark. E-mail: mette.bach.larsen@alm.au.dk

ACCEPTED: 27 November 2013.

CONFLICTS OF INTEREST: Disclosure forms provided by the authors are available with the full text of this article at www.danmedj.dk.

LITERATURE

1. Engholm G, Ferlay J, Gjerstorff M et al. NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 4.0. Association of the Nordic Cancer Registries. Copenhagen: Danish Cancer Society, 2011.
2. The Danish National Board of Health. Dødsårsagsregisteret 2010. Copenhagen: Danish National Board of Health, 2008.
3. Sørensen HT, Lash TL. Use of administrative hospital registry data and a civil registry to measure survival and other outcomes after cancer. Clin Epidemiol 2011;3(suppl 1):1-2.
4. Hansen RP, Vedsted P, Sokolowski I et al. Time intervals from first symptom to treatment of cancer: a cohort study of 2,212 newly diagnosed cancer patients. BMC Health Serv Res 2011;11:284.
5. The Danish National Board of Health. Fællesindhold for basisregistrering af sygehuspatienter. www.sst.dk/Webudgivelser/FaellesIndhold/Forside.aspx (21 Dec 2011).
6. Sørensen HT, Christensen T, Schlosser HK et al. Use of medical databases in clinical epidemiology. Aarhus: Department of Clinical Epidemiology, Aarhus University Hospital, 2008.
7. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scand J Public Health 2011;39:30-3.
8. The Danish National Board of Health. Det moderniserede Cancer Register. Copenhagen: The Danish National Board of Health, 2009.
9. Gjerstorff ML. The Danish Cancer Registry. Scand J Public Health 2011;39:42-5.
10. Hansen RP. Delay in the diagnosis of cancer, 1 ed. Aarhus: Faculty of Health Sciences, University of Aarhus, 2008.
11. Bjerager M. Delay in diagnosis and treatment of lung cancer, 1 ed. Aarhus: Research Unit and Department of General Practice, Faculty of Health Sciences, University of Aarhus, 2006.
12. The Danish National Board of Health. Evaluering af landspatientregisteret 1990. Copenhagen: The Danish National Board of Health, 1993.
13. Lidegaard O, Hammerum MS. The National Patient Registry as a tool for continuous production and quality control. Ugeskr Læger 2002;164: 4420-3.
14. Lidegaard O, Vestergaard CH, Hammerum MS. Quality monitoring based on data from the Danish National Patient Registry. Ugeskr Læger 2009;171:412-5.
15. Nørgaard M, Skriver MV, Gregersen H et al. The data quality of haematological malignancy ICD-10 diagnoses in a population-based hospital discharge registry. Eur J Cancer Prev 2005;14:201-6.
16. Helqvist L, Erichsen R, Gammelager H et al. Quality of ICD-10 colorectal cancer diagnosis codes in the Danish National Registry of Patients. Eur J Cancer Care 2012;21:722-7.
17. Gammelager H, Christiansen CF, Johansen MB et al. Quality of urological cancer diagnoses in the Danish National Registry of Patients. Eur J Cancer Prev 2012;21:545-51.
18. Vedsted P, Hansen RP, Bro F. General practice and early cancer diagnosis. Ugeskr Læger 2011;173:1712-5.
19. Black N, Payne M. Directory of clinical databases: improving and promoting their use. Qual Saf Health Care 2003;12:348-52.
20. Baron JA, Weiderpass E. An introduction to epidemiological research with medical databases. Ann Epidemiol 2000;10:200-4.