CURRENT ONCOLOGY
A Canadian Cancer Research Journal

SHORT COMMUNICATION

Validation in Alberta of an administrative data algorithm to identify cancer recurrence

Z.F. Cairncross MPH,* G. Nelson MD PhD,* L. Shack PhD,† and A. Metcalfe PhD*

ABSTRACT

Background Readily available population-based data about cancer recurrence would improve surveillance and research for women of reproductive age.

Methods We randomly selected 200 women from the Alberta Cancer Registry who had received a cancer diagnosis and who ever had a pregnancy between 2003 and 2012. Administrative data were obtained and linked. Several definitions of recurrence were assessed using various minimum lengths of time between the initial diagnosis date and subsequent diagnoses or treatments, or both. Chart review was used as a "gold standard" definition of recurrence.

Results Chart review identified recurrences in 26 women. The definition that best captured "recurrence" was 2 or more cancer diagnosis codes 10 or more months from the diagnosis date [sensitivity: 80.8%; 95% confidence interval (CI): 60.7% to 93.5%; specificity: 81.0%; 95% CI: 74.4% to 86.6%; positive predictive value: 38.9%; 95% CI: 25.9% to 53.1%; negative predictive value: 96.6%; 95% CI: 92.2% to 98.9%; kappa = 0.42; 95% CI: 0.28 to 0.57].

Conclusions Recurrence in reproductive-aged women can be captured with moderate validity using administrative data, but should be interpreted with caution.

Key Words Recurrence, breast cancer, validation, administrative data, registries

Curr Oncol. 2020 June:27(3)e343–e346 www.current-oncology.com

INTRODUCTION

Cancer recurrence, risk of death, and disease-free survival are common endpoints used in cancer studies to evaluate the effect that cancer might have on an individual's long-term health[1]. The use of population-based data, including cancer registries and administrative databases, has become common in the investigation of cancers. By investigating all applicable cases over a long study period or large geographic area (that is, population-based retrospective cohorts), researchers can obtain a wealth of information at a relatively small cost[2]. However, cancer registries cannot typically record recurrence for a number of reasons, including cost, volume of cases, and movement of patients[3,4], leaving researchers with three options:

- Review personal health records, which have higher validity and reliability, but can be costly and inefficient for large sample sizes[2].
- Apply algorithms to administrative data to identify evidence of recurrence, a method that has been shown to have variable levels of validity[5–8].
- Forgo the investigation of recurrence, a choice that critically affects patients and providers.

In the case of cancers in women of reproductive age, the risk of recurrence is an important consideration in how an affected woman might plan future pregnancies after a cancer diagnosis[9]—that is, within what timeframe would she be most likely to have a cancer-free pregnancy? With cancer being the 2nd leading cause of death in women of reproductive age and there being a long time for potential recurrence, reaching a better understanding of recurrence becomes increasingly important[10].

In the present study, we assessed the ability of administrative data, compared with personal health records, to accurately ascertain cancer recurrence in women of reproductive age.

Correspondence to: Amy Metcalfe, Department of Obstetrics and Gynecology, University of Calgary, 1403 29 Street NW, Calgary, Alberta T2N 2T9.
E-mail: amy.metcalfe@albertahealthservices.ca **DOI:** https://doi.org/10.3747/co.27.5861

METHODS

Study Population

Individual-level data were obtained from the Alberta Cancer Registry (ACR). The study population consisted of women 18–50 years of age from Calgary, Alberta, who had a cancer diagnosis (excluding non-melanoma skin cancer and cervical carcinoma *in situ)* and who ever had a pregnancy reaching at least 20 weeks' gestation between 1 January 2003 and 31 December 2012. The study sample consisted of 200 women randomly selected from the cohort meeting those conditions. All women included in the analysis had a minimum of 5 years of follow-up after their initial cancer diagnosis. Ethics approval for the study was granted by the Health Research Ethics Board of the Alberta Cancer Committee.

Data Sources

The ACR is a population-based registry that collects detailed information for all new cancer cases in the province of Alberta; it has received gold star certification from the North American Association of Central Cancer Registries, with more than 95% case ascertainment[11]. New cases are identified by, and forwarded from, physicians and laboratories in the province, from visits to cancer facilities, and from electronic linkage to Alberta Vital Statistics for information about deaths[12]. Variables collected include cancer site, morphology, cancer stage, patient demographics, and diagnosis date.

The National Ambulatory Care Reporting System (NACRS) contains data on all emergency department visits, day surgeries, and some hospital- and community-based outpatient clinics. The NACRS is available in 8 of the 13 Canadian provinces and territories[13]. The data are abstracted from hospital charts by trained medical records personnel and include admission and discharge dates, and up to 25 diagnoses and 15 procedures for each episode of care[13]. Diagnoses are coded using the Canadian modification to the *International Statistical Classification of Diseases and Related Health Problems,* 10th Revision[14]. Procedures are coded using the *Canadian Classification of Health Interventions*[15].

Personal health records are collected and maintained by Alberta Health Services and contain information about an individual's interactions with the health care system, such as reports from health care providers, treatment information, and test results.

Defining Cancer Recurrence

Algorithms for identifying cancer recurrence were based on previous literature that used cancer registries and administrative data holdings for coding recurrence based on various combinations of treatment dates and diagnosis or procedure codes[4,7]. The definitions were then adapted for the data holdings in Alberta by the study analyst in consultation with a gynecologic oncologist. Recurrence was identified in NACRS by measuring the time from the diagnosis date (as recorded in the ACR) to either subsequent cancer diagnosis codes or subsequent cancer diagnosis codes plus a procedure code for a physical or physiologic therapeutic intervention (Section 1 of the *Canadian Classification of Health Interventions)*. We tested definitions ranging from 6 to 18 months between the diagnosis date and any subsequent cancer diagnosis codes to identify three definitions: high sensitivity, high specificity, and high sensitivity and specificity. Presence or absence of cancer recurrence was assigned to each patient once for each case definition.

Chart Review

Patient charts were considered the reference standard for all analyses. Patient charts, beginning after the diagnosis date, were reviewed by a gynecologic oncologist blinded to the potential recurrence status to determine if a woman experienced a cancer recurrence. Cancer recurrence was documented if a progress note or treatment plan in the chart clearly indicated a recurrence. A standardized data extraction sheet was used to record whether the cancer had recurred, and if it had, the first date of the recurrence.

Statistical Analyses

Proportions and 95% confidence intervals (CIs) are used to describe relevant characteristics of the study population, including age, cancer site, and cancer stage at diagnosis. To assess the validity of NACRS compared with patient charts, we calculated sensitivity, specificity, positive and negative predictive values, and kappa coefficients for each definition across all cancer types. A stratified analysis was also conducted for breast cancer, the most prevalent cancer type. In the primary analysis, patients without available charts were assumed not to have experienced recurrence. A sensitivity analysis was conducted restricted to the subset of patients who had a record in the ACR and who had at least one chart to review. Additional sensitivity analyses considered subsequent cancer diagnoses only when cancer was responsible for most of the services used.

RESULTS

Table I presents descriptive characteristics for the sample. Mean age of the patients was 33.8 ± 6.6 years. The most prevalent cancers in the sample were breast (32.5%, n = 65), thyroid (24.0%, n = 48), and melanoma of skin (11.0%,

TABLE I Descriptive characteristics of study population

Characteristic	Value	95% CI
Mean age at Dx (years)	33.8±6.6	
Cancer site [*n* (%)]		
Breast	65 (32.5)	26.1 to 39.5
Thyroid	48 (24.0)	18.3 to 30.5
Melanoma of skin	22 (11.0)	7.0 to 16.2
Hodgkin lymphoma	12 (6.0)	3.1 to 10.3
Ovary	9 (4.5)	2.1 to 8.4
Other	44 (22.0)	16.5 to 28.4
Cancer stage at Dx [*n* (%)]		
I	52 (26.0)	20.1 to 32.7
II	43 (21.5)	16.0 to 27.9
III	24 (12.0)	7.8 to 17.3
IV	8 (4.0)	1.7 to 7.7
Unknown or unstageable	73 (36.5)	29.8 to 43.6

CI = confidence interval; Dx = diagnosis.

$n = 22$). Cancers were diagnosed at stage I (26.0%, $n = 52$), stage II (21.5%, $n = 43$), stage III (12.0%, $n = 24$), and stage IV (4.0%, $n = 8$). Approximately 37% of the cancers ($n = 73$) had an unknown stage or were unstageable. Chart review revealed recurrences in 26 women (13.0%).

The validity of administrative data in determining cancer recurrence varied with the case definitions (Table II). The definition that best optimized sensitivity and specificity was 2 or more cancer diagnosis codes 10 or more months from the diagnosis date (sensitivity: 80.8%; 95% CI: 60.7% to 93.5%; specificity: 81.0%; 95% CI: 74.4% to 86.6%; positive predictive value: 38.9%; 95% CI: 25.9% to 53.1%; negative predictive value: 96.6%; 95% CI: 92.2% to 98.9%; kappa = 0.42; 95% CI: 0.28 to 0.57). Results for patients with complete data and with breast cancer were very similar to results in the main analysis. Low validity was found for definitions that considered subsequent diagnoses only when cancer was responsible for most of the services used.

DISCUSSION AND CONCLUSIONS

We randomly selected and reviewed the charts of 200 women of reproductive age included in the ACR to determine the validity of multiple case definitions for identifying cancer recurrences. The results of the study are encouraging for future population-based research into cancers in women of reproductive age. Of the case definitions assessed, the definition using a minimum of 10 months between the diagnosis date and 2 or more subsequent cancer diagnosis codes had the best overall validity. That definition can be used in Alberta for women of reproductive age, but should be interpreted with caution; if used in studies outside Alberta, the definition should be re-validated before use.

Results of this study are unsurprising, given the mixed results in the existing literature where similar methods were used[4]. A review of the measurement of cancer recurrence based on administrative data in the United States found that the success of algorithms is often encouraging in small single-site studies[5,6], but can encounter high rates of misclassification in larger population-based studies[7,8]. To account for how health data are coded differently in different jurisdictions, algorithms identifying recurrence in administrative data must therefore be validated in every health system before use. The objective of the present study was to test whether an algorithm could be created to identify any recurrence in women of reproductive age; the creation of multiple definitions specific to various cancer sites was therefore outside the scope and resources of the study. However, existing literature shows that the validity of definitions can often be improved if based on site-specific diagnoses, procedures, treatments, and recurrence patterns[5,7].

Our study has several limitations. In Alberta, cancer data are recorded largely in patient charts, not administrative data. Although we were able to use NACRS to develop moderately accurate definitions, the data lack detail. For example, cancer diagnoses recorded 1 year after the initial diagnosis date might reflect recurrence, progression, or refractory cancer, or might simply have been an indication in a chart that a patient previously had cancer. To account for the latter situation, we conducted a sensitivity analysis using only most-responsible diagnosis codes, but the result

TABLE II Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and kappa of the National Ambulatory Care Reporting System for cancer recurrence

Recurrence definition	Sensitivity		Specificity		PPV		NPV		Kappa	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
All cancer sites (n=*200*)										
High sensitivity[a]	96.2	80.4 to 99.9	67.8	60.3 to 74.7	30.9	21.1 to 42.1	99.2	95.4 to 100	0.34	0.23 to 0.45
High specificity[b]	50.0	29.9 to 70.1	93.1	88.3 to 96.4	52.0	31.3 to 72.2	92.6	87.6 to 96.0	0.44	0.25 to 0.62
High sensitivity and specificity[c]	80.8	60.7 to 93.5	81.0	74.4 to 86.6	38.9	25.9 to 53.1	96.6	92.2 to 98.9	0.42	0.28 to 0.57
All cancer sites, complete data (n=*169*)										
High sensitivity[a]	96.2	80.4 to 99.9	66.4	58.1 to 74.1	34.3	23.5 to 46.3	99.0	94.3 to 100	0.36	0.24 to 0.48
High specificity[b]	50.0	29.9 to 70.1	91.6	85.8 to 95.6	52.0	31.3 to 72.2	91.0	85.1 to 95.1	0.42	0.23 to 0.61
High sensitivity and specificity[c]	80.8	60.7 to 93.5	81.1	73.7 to 87.2	43.8	29.5 to 58.8	95.9	90.6 to 98.6	0.46	0.31 to 0.61
Breast cancer (n=*65*)										
High sensitivity[a]	100	80.5 to 100	66.7	51.6 to 79.6	51.5	33.5 to 69.2	100	89.1 to 100	0.51	0.33 to 0.69
High specificity[b]	52.9	27.8 to 77.0	87.5	74.8 to 95.3	60.0	32.3 to 83.7	84.0	70.9 to 92.8	0.42	0.17 to 0.67
High sensitivity and specificity[c]	88.2	63.6 to 98.5	85.4	72.2 to 93.9	68.2	45.1 to 86.1	95.4	84.2 to 99.4	0.67	0.48 to 0.87

[a] One or more cancer diagnosis codes 10 or more months from diagnosis.
[b] One or more cancer diagnosis codes and a treatment procedure code 9 or more months from diagnosis.
[c] Two or more cancer diagnosis codes 10 or more months from diagnosis.
CI = confidence interval.

was a decrease in validity. The study was limited to women of reproductive age, and so results are not generalizable to the larger population, given different treatment and recurrence patterns. Nevertheless, the available literature about algorithms to detect recurrence use many of the same overarching methods, and therefore can and should be explored to expand research about recurrence across various age groups. Because of a small sample size, we were unable to test definitions in cancers other than breast cancer.

Risk of recurrence is a critical outcome in cancer epidemiology, a field that is increasingly using population-based methods. Lack of information about recurrence precludes evidence-based discussions between physicians and affected patients, especially patients with rare cancers in which population-based surveillance systems are the primary mechanism to evaluate the natural history of the disease and the effect of various treatment strategies on clinically relevant outcomes such as survival and recurrence.

ACKNOWLEDGMENTS

Funding for this study was provided by Canadian Institutes of Health Research.

CONFLICT OF INTEREST DISCLOSURES

We have read and understood *Current Oncology*'s policy on disclosing conflicts of interest, and we declare that we have none.

AUTHOR AFFILIATIONS

*Department of Obstetrics and Gynecology, University of Calgary, and †Cancer Research and Analytics, CancerControl Alberta, Alberta Health Services, Calgary, AB.

REFERENCES

1. Lee SJ, Earle CC, Weeks JC. Outcomes research in oncology: history, conceptual framework, and trends in the literature. *J Natl Cancer Inst* 2000;92:195–204.
2. Thygesen LC, Ersbøll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. *Eur J Epidemiol* 2014;29:551–8.
3. In H, Bilimoria KY, Stewart AK, *et al.* Cancer recurrence: an important but missing variable in national cancer registries. *Ann Surg Oncol* 2014;21:1520–9.
4. Warren JL, Yabroff KR. Challenges and opportunities in measuring cancer recurrence in the United States. *J Natl Cancer Inst* 2015;107:djv134.
5. Deshpande AD, Schootman M, Mayer A. Development of a claims-based algorithm to identify colorectal cancer recurrence. *Ann Epidemiol* 2015;25:297–300.
6. Earle CC, Nattinger AB, Potosky AL, *et al.* Identifying cancer relapse using SEER–Medicare data. *Med Care* 2002;40(suppl):IV75–81.
7. Chubak J, Yu O, Pocobelli G, *et al.* Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *J Natl Cancer Inst* 2012;104:931–40.
8. Hassett MJ, Ritzwoller DP, Taback N, *et al.* Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using two large contemporary cohorts. *Med Care* 2014;52:e65–73.
9. Partridge AH, Ruddy KJ. Fertility and adjuvant treatment in young women with breast cancer. *Breast* 2007;16(suppl2):175–81.
10. Lee YY, Roberts CL, Dobbins T, *et al.* Incidence and outcomes of pregnancy-associated cancer in Australia, 1994–2008: a population-based linkage study. *BJOG* 2012;119:1572–82.
11. Fisher S, Gao H, Yasui Y, Dabbs K, Winget M. Treatment variation in patients diagnosed with early stage breast cancer in Alberta from 2002 to 2010: a population-based study. *BMC Health Serv Res* 2015;15:35.
12. CancerControl Alberta. *The 2019 Report on Cancer Statistics in Alberta*. Edmonton, AB: Alberta Health Services; 2019.
13. Canadian Institute for Health Information (CIHI). *Data Quality Documentation, National Ambulatory Care Reporting System—Current-Year Information, 2017–2018.* Ottawa, ON: CIHI; 2018.
14. Canadian Institute for Health Information (CIHI). *International Statistical Classification of Diseases and Related Health Problems. Tenth Revision, Canada.* Ottawa, ON: CIHI; 2009.
15. Canadian Institute for Health Information (CIHI). *Canadian Classification of Health Interventions.* Ottawa, ON: CIHI; 2009.