Received: 6 November 2017 | Revised: 8 June 2018 | Accepted: 21 July 2018

DOI: 10.1002/pds.4641

Identification of incident pancreatic cancer in Ontario administrative health data: A validation study

Jennifer W. Wu[1,2] | Laurent Azoulay[1,2,3] | Anjie Huang[4] | Michael Paterson[4,5,6] | Fangyun Wu[4] | Matthew H. Secrest[2] | Kristian B. Filion[1,2,7]

[1] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

[2] Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal, Quebec, Canada

[3] Gerald Bronfman Department of Oncology, McGill University, Montreal, Quebec, Canada

[4] Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada

[5] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

[6] Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada

[7] Division of Clinical Epidemiology, Department of Medicine, McGill University, Montreal, Quebec, Canada

Correspondence

K. B. Filion, Departments of Medicine and of Epidemiology, Biostatistics, and Occupational Health, Jewish General Hospital/McGill University, 3755 Cote Ste-Catherine Road, Suite H410.1, Montreal, Quebec H3T 1E2, Canada.

Email: kristian.filion@mcgill.ca

Funding information

Fonds de recherche du Québec—Santé; Canadian Institutes of Health Research, Grant/Award Number: DSE-146021; ICES

Abstract

Purpose: To validate three approaches for identifying incident cases of pancreatic cancer in Ontario administrative claims data.

Methods: We created a cohort using Ontario (Canada) administrative health data from 2002 to 2012 and identified cases of pancreatic cancer with three approaches, using the Ontario Cancer Registry (OCR) as the reference standard. In the *any diagnosis* approach, cases were defined by primary or secondary diagnostic codes for pancreatic cancer in outpatient or inpatient records. In the *any inpatient diagnosis* approach, cases were defined using only diagnoses in hospital discharge abstracts. In the *algorithm* approach, cases were identified by an algorithm that combined the first two approaches. Comparing each approach to the OCR, we calculated the expected value and 95% confidence interval (CI) of the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We also compared the event dates using each approach with those recorded in the OCR.

Results: Among a total of 12 060 837 patients in Ontario administrative health data sources, 13 999 incident pancreatic cancer cases were identified in the OCR. Sensitivity ranged from 72.5% (*algorithm*) to 97.5% (*any diagnosis*), and PPV ranged from 38.4% (*any diagnosis*) to 78.9% (*any inpatient diagnosis*). Specificity and NPV were ~100% for all approaches. The median absolute difference in cancer event date ranged 0 to 15 days. The *any inpatient diagnosis* method had the highest PPV (78.9%; 95% CI: 78.2-79.5%) and moderate sensitivity (86.6%; 95% CI: 86.0-87.2%).

Conclusion: Inpatient diagnoses of pancreatic cancer in Ontario administrative heath data are suitable for pancreatic cancer case identification.

KEYWORDS

administrative health databases, cancer registry, pancreatic cancer, pharmacoepidemiology, validation

1 | INTRODUCTION

Pancreatic cancer is the fourth leading cause of cancer-related deaths in Canada and the United States, with a case-fatality rate of 90% over

This provided manuscript is the sole work of the identified contributing authors. It has been accepted for poster presentation at the 2018 International Conference on Pharmacoepidemiology and Therapeutic Risk Management in August 2018.

5 years.[1,2] Observational epidemiologic research is greatly needed to help mitigate the high burden of disease and mortality attributable to pancreatic cancer—for example, by identifying or ruling out risk factors,[3] assessing treatment effectiveness among patients,[4] and tracking population health statistics.[5] Such observational studies require efficient and accurate identification of incident pancreatic cancer cases in secondary data sources.

Cancer registries represent the preferred data sources for the study of incident pancreatic cancer cases and cancer-related deaths. Although provincial cancer registries are mandated in Canada, these registries are often less readily accessible for research purposes than administrative health data sources,[6] due to time lags for updates, and in some cases, longer data access delays. Several cancer case definitions have been previously validated in administrative data,[7-16] but these have focused primarily on screened cancers such as breast, colorectal, and prostate; the generalizability of these algorithms to non-screened cancers is unclear. Few algorithms have been developed for the identification of incident pancreatic cancer[11,14-16]; these previous studies were conducted in the US (Medicare), the United Kingdom, Australia, and Nordic countries. Because previous studies suggested the generalizability of such validation studies across health and data systems was limited,[17] there remains a need to examine the validity of such algorithms using Canadian administrative databases. Furthermore, previous studies have largely focused on the validity of hospital discharge data[14,15] or general practitioner data,[17] or were restricted to the elderly,[11] with little information available regarding validity in linked, population-based databases.

Therefore, the objective of this study was to determine the validity of three different approaches for identifying incident pancreatic cancer cases in administrative health data from Ontario, Canada's most populous province, using the Ontario Cancer Registry (OCR) as the reference standard data source.

2 | METHODS

2.1 | Data sources

Several Ontario administrative health databases were used in this study: the Ontario Health Insurance Plan Claims Database (OHIP), which captures claims for all inpatient and outpatient physician services[18,19]; the Canadian Institute of Health Information Discharge Abstract Database, which contains diagnoses and procedures associated with all acute hospital admissions; the Canadian Institute for Health Information National Ambulatory Care Reporting System, which contains records for all emergency department and outpatient cancer clinic encounters; the Institute for Clinical Evaluative Sciences Physician Database,[18,19] a registry of the characteristics of all physicians and surgeons practicing in Ontario; and the OHIP Registered Persons Database, the provincial health insurance registry. The OCR served as our reference standard data source for the identification of patients with incident pancreatic cancer. These databases are held securely in linked coded form and analyzed at the Institute for Clinical Evaluative Sciences (ICES, www.ices.on.ca).

Key Points

- Few validation studies have evaluated incident pancreatic cancer case definitions in administrative health databases compared with cancer registries, considered the reference standard.
- Three approaches for identifying pancreatic cancer cases were evaluated using (1) inpatient diagnoses only; (2) any inpatient or outpatient diagnosis; and (3) an algorithm that combined 1 and 2.
- The optimal approach was to use inpatient diagnoses only, which yielded the highest positive predictive value, had moderate sensitivity, and did not appreciably affect the date of cancer diagnosis compared with that recorded in the cancer registry (median difference = 0 days).
- In the absence of cancer registry data, future epidemiologic studies could use inpatient diagnosis codes to identify incident pancreatic cancer.

Since 1964, the OCR has collected information on Ontario residents with new cancer diagnoses or cancer-related causes of death.[20] All cancers are included except for non-melanoma skin cancer. There are a multitude of purposes to the OCR, including research and surveillance. The OCR depends on four major data sources: hospital discharge abstracts, pathology reports, regional cancer center data, and death certificates.[21] The OCR includes two components: the decommissioned Ontario Cancer Registry Information System (1964 to 2009), which followed International Agency for Research on Cancer rules, and the current OCR (2010 to present), which follows the National Cancer Institute's Surveillance, Epidemiology and End Results multiple primary and histology coding rules.[22] Previous studies have found that the OCR is 98.5% complete,[23] with a capture rate > 95% for pancreatic cancer.[24] Cases are not medically validated; rather, the registry uses a combination of deterministic and probabilistic linkage to establish a composite record from all the data sources on new cancers and cancer-related deaths.

2.2 | Study design and population

Using the Ontario administrative health databases, we created a cohort of patients aged 18 years or older at any time between 1 April 2002 and 31 December 2012 who had at least 5 years of OHIP coverage before they entered the cohort. For each patient, the date of cohort entry was defined by the latest of their 18th birthday, the date on which they had 5 years of OHIP coverage, and the beginning of the study period (1 April 2002). We excluded patients with a history of pancreatic cancer as documented in the OCR any time on or before cohort entry, patients with a history of pancreatectomy in the preceding 5 years, and those who died on or before the cohort entry date.

We followed patients until an incident pancreatic cancer event (see details below), death, loss of OHIP coverage, or the end of the study period (31 December 2012).

2.3 | Approaches to identifying patients with pancreatic cancer

Three approaches were used to identify incident, primary pancreatic cancer in Ontario administrative health databases. In all approaches, pancreatic cancer was defined by International Classification of Diseases (ICD) 9th and 10th Revision codes (ICD-9: 157.0-157.9; ICD-10: C25.x). In the first approach, the *any diagnosis* approach, pancreatic cancer cases were defined by the presence of either an inpatient primary or secondary diagnosis code on a hospital discharge abstract or an outpatient diagnosis on a physician service claim or emergency department record. For this method, the date of pancreatic cancer diagnosis was the date of the first such encounter, using the date of admission for the inpatient record and the date of service for the outpatient records. In the second approach, the *any inpatient diagnosis* approach, pancreatic cancer cases were limited to those identified in hospital discharge abstracts, with the admission date serving as the date of pancreatic cancer diagnosis. Lastly, we created an algorithm that combined these approaches (Figure 1). In brief, this *algorithm* approach selected the subset of pancreatic cancer cases identified through the first two approaches for which a confirmatory event was available within 90 days of the diagnosis date. Confirmatory events were defined by a hospital discharge abstract with a primary or secondary diagnosis code for pancreatic cancer or a physician service claim submitted by a medical or radiation oncologist.

2.4 | Patient characteristics

We described and compared the characteristics of the pancreatic cancer patients we identified using our three approaches and the OCR. Characteristics included age at diagnosis, sex, year of cohort entry, number of hospitalizations in the 365 days preceding diagnosis, presence of chronic pancreatitis during the year preceding diagnosis, and Deyo-Charlson Comorbidity Score[25] based upon hospitalizations over the 3 years preceding diagnosis.

2.5 | Data analysis

The accuracy of our three approaches was assessed by sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).[26] The corresponding 95% confidence intervals (CIs) were calculated based on the exact binomial distribution. The OCR served as our reference standard. We also calculated the absolute difference in date of cancer diagnosis between each of our administrative health database approaches and the OCR for patients mutually identified by the approach and the OCR. Because many pharmacoepidemiologic studies are conducted in elderly patients, we did a sensitivity analysis in a sub-cohort of patients 65 years or older from 1 April 2002 to 31 December 2012. We applied the same cohort exclusion criteria as described previously. In sensitivity analyses conducted to examine the potential presence of temporal trends, we repeated our primary analysis stratified by calendar year.

All analyses were performed at ICES using SAS (version 9.4, Cary, NC).

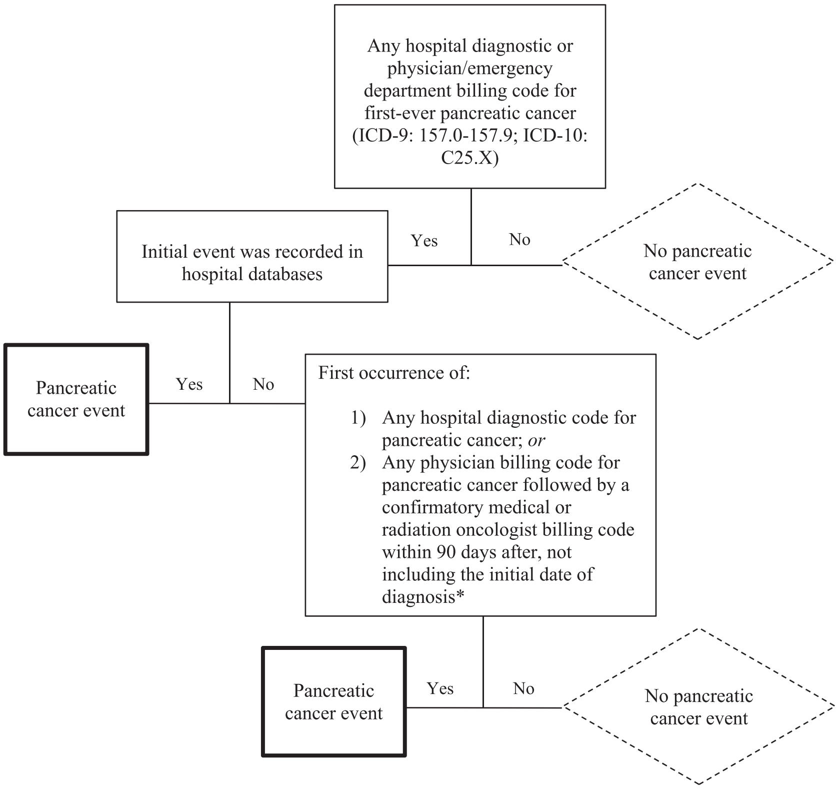
Any hospital diagnostic or physician/emergency department billing code for first-ever pancreatic cancer (ICD-9: 157.0-157.9; ICD-10: C25.X)
Yes
No
Initial event was recorded in hospital databases
No pancreatic cancer event
Pancreatic cancer event
Yes
No
First occurrence of:
1) Any hospital diagnostic code for pancreatic cancer; or
2) Any physician billing code for pancreatic cancer followed by a confirmatory medical or radiation oncologist billing code within 90 days after, not including the initial date of diagnosis*
Pancreatic cancer event
Yes
No
No pancreatic cancer event

FIGURE 1 Algorithm to identify incident pancreatic cancer in Ontario's administrative health databases based on outpatient and inpatient billing codes. Abbreviations: ICD-9/10, International Classification of Disease, 9th/10th revision. * If a patient had more than one confirmatory event during the 90-day window, the earliest event determined the index date

3 | RESULTS

A total of 12 060 837 patients were identified in Ontario administrative health databases after applying our exclusion criteria (Figure 2). We identified 35 522 pancreatic cancer cases in the *any diagnosis* approach, 15 367 in the *any inpatient diagnosis* approach, 13 610 in the *algorithm* approach, and 13 999 in the OCR. The *any inpatient diagnosis* approach and OCR were more similar in terms of patient characteristics compared with the other two approaches (Table 1). In general, cases identified through the *any inpatient diagnosis* approach and the OCR were slightly older, had fewer hospitalizations, and had fewer comorbidities compared with the other two approaches. All three of our administrative database approaches identified more male cases than the OCR.

The *any inpatient diagnosis* approach had the highest PPV (78.9%; Table 2), and it also had the second highest sensitivity (86.6%). The *algorithm* approach yielded slightly lower PPV (74.6%) and sensitivity (72.5%). The *any diagnosis* approach had the lowest PPV (38.4%) and highest sensitivity (97.5%). Specificity and NPV were ~100.0% for all approaches.

The mean difference in timing of pancreatic cancer events between the administrative data approaches and the OCR ranged from 43 to 55 days (Table 3), whereas the median difference ranged from 0 to 15 days. The *any inpatient diagnosis* approach had the lowest median difference in timing of pancreatic cancer events (0 days). The *any diagnosis* approach had a similar median difference in timing of pancreatic cancer events (2 days). In contrast, the *algorithm* approach had the highest median difference in timing of pancreatic cancer events (15 days).

From the study population, we identified a total of 2 842 226 patients aged 65 years or older (Appendix 1). In this population, the *any inpatient diagnosis* approach was the most similar to the OCR in terms of patient characteristics (Appendix 2). Consistent with the primary analysis in the overall population, in the subpopulation of elderly patients, the *any inpatient diagnosis* approach had the highest PPV (76.1%), the second highest sensitivity (86.8%) (Appendix 3), and the lowest median difference in timing of pancreatic cancer events (0 days) (Appendix 4) relative to OCR.

Sensitivity analyses stratified by calendar year revealed some variability across years. However, general patterns were consistent over time, suggesting that any temporal trends that are present would not impact the conclusions drawn regarding the validity of the approaches assessed (Appendix 5).

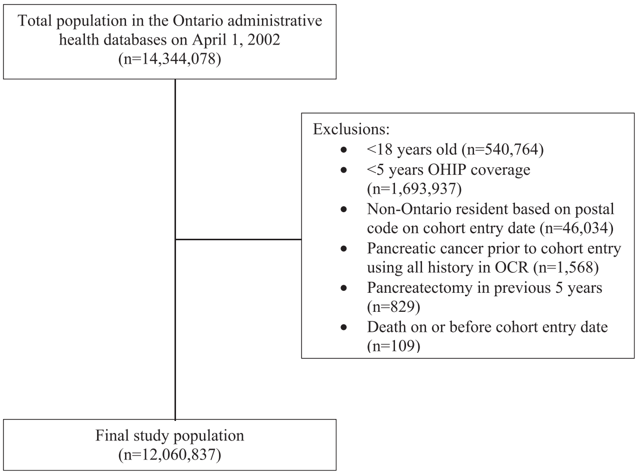
Total population in the Ontario administrative health databases on April 1, 2002 (n=14,344,078)
Exclusions:
<18 years old (n=540,764)
<5 years OHIP coverage (n=1,693,937)
Non-Ontario resident based on postal code on cohort entry date (n=46,034)
Pancreatic cancer prior to cohort entry using all history in OCR (n=1,568)
Pancreatectomy in previous 5 years (n=829)
Death on or before cohort entry date (n=109)
Final study population (n=12,060,837)

FIGURE 2 Flow chart describing selection of patients from Ontario's administrative health databases. Abbreviations: OCR, Ontario Cancer Registry; OHIP, Ontario Health Insurance Plan Claims Database

4 | DISCUSSION

We evaluated three approaches for identifying patients with incident pancreatic cancer in Ontario administrative health databases, using the OCR as the reference standard. While the approaches had tradeoffs that may recommend their use differentially according to the research question of interest,[26] we argue that the *any inpatient diagnosis* approach is the most suitable for epidemiologic research due to its simplicity and relatively high PPV (78.9%) and sensitivity (86.6%). The *any inpatient diagnosis* approach was also the least likely to introduce gaps in the estimated timing of cancer diagnoses, with a median difference in diagnosis timing of zero days when compared with the OCR. Finally, the *any inpatient diagnosis* method also performed better than the other approaches in older patients, suggesting it also is preferable for the study of seniors.

The superior PPV of the *any inpatient diagnosis* approach when compared with the other approaches likely relates to the logistics of pancreatic cancer diagnosis. In current practice, pancreatic cancer is more likely to be diagnosed and treated in-hospital. In the United Kingdom, for instance, 50% of incident pancreatic cancer cases are identified in the emergency room.[27] This is largely due to the pathophysiology of pancreatic cancer, which includes symptoms of abdominal pain and jaundice.[28] Consequently, patients are referred to the hospital, where diagnoses are based on reliable methods such as ultrasounds and computed tomography scans. Because incident pancreatic cancer shares symptoms with other diseases such as pancreatitis,[28] outpatient claims may contain a large number of false positives, where patients without pancreatic cancer are diagnosed with pancreatic cancer.[29] This could explain the low PPV of the *any diagnosis* approach. We note that the high NPV of the *any inpatient diagnosis* approach and all other approaches relates to the low prevalence of pancreatic cancer in our source population.[30]

While the *any inpatient diagnosis* approach had the best PPV, it had only moderately high sensitivity when compared with the *any diagnosis* method. This results from the hierarchical nature of these methods, ie, all diagnoses captured by the *any inpatient diagnosis* approach were also captured by the *any diagnosis* approach. The requirement for the *algorithm* approach of a confirmatory event likely decreased its sensitivity through exclusion of some true cancer patients for whom confirmatory events were unavailable.

Our conclusion that the *any inpatient diagnosis* approach was the most valid for health care research was consistent with previous validation studies of cancer. Previous pancreatic cancer case definitions based upon hospital discharge abstracts or hospital databases have yielded moderate to high sensitivities (range: 80.2-95%)[11,14,16] and

TABLE 1 Characteristics of the pancreatic cancers cases identified from the three approaches using Ontario's administrative health databases and the Ontario Cancer Registry[a]

	Approach[b]			
Characteristics	Any Diagnosis *N* = 35 522	Any Inpatient Diagnosis *N* = 15 367	Algorithm *N* = 13 610	Ontario Cancer Registry *N* = 13 999
Age (years)	67.8 ± 14.4	70.8 ± 12.5	68.6 ± 12.0	69.8 ± 12.5
18-34	828 (2.3)	85 (0.6)	76 (0.6)	75 (0.5)
35-44	1627 (4.6)	308 (2.0)	315 (2.3)	315 (2.3)
45-54	3941 (11.1)	1283 (8.3)	1388 (10.2)	1315 (9.4)
55-64	6919 (19.5)	2880 (18.7)	3066 (22.5)	2853 (20.4)
65-74	9069 (25.5)	4221 (27.5)	4049 (29.8)	3913 (28.0)
75-84	9403 (26.5)	4623 (30.1)	3661 (26.9)	3984 (28.5)
85-94	3518 (9.9)	1854 (12.1)	1016 (7.5)	1442 (10.3)
95+	217 (0.6)	113 (0.7)	39 (0.3)	102 (0.7)
Male	18 004 (50.7)	7703 (50.1)	7101 (52.2)	6899 (49.3)
Calendar year				
2002-2004	9105 (25.6)	3444 (22.4)	2771 (20.4)	3074 (22.0)
2005-2007	8977 (25.3)	4078 (26.5)	3470 (25.5)	3787 (27.1)
2008-2010	9914 (27.9)	4442 (28.9)	4143 (30.4)	4089 (29.2)
2011-2012	7526 (21.2)	3403 (22.1)	3226 (23.7)	3049 (21.8)
Number of hospitalizations in preceding 365 days	0.7 ± 1.1	0.5 ± 0.9	0.8 ± 1.0	0.4 ± 0.8
0	20 756 (58.4)	10 512 (68.4)	6874 (50.5)	10 506 (75.0)
1	9278 (26.1)	3306 (21.5)	4520 (33.2)	2525 (18.0)
2	3352 (9.4)	1012 (6.6)	1474 (10.8)	660 (4.7)
3+	2136 (6.0)	537 (3.5)	742 (5.5)	308 (2.2)
Chronic pancreatitis	689 (1.9)	185 (1.2)	184 (1.4)	155 (1.1)
Charlson comorbidity score[c]	2.3 ± 2.5	1.8 ± 2.2	3.1 ± 2.7	1.2 ± 1.6
0	6227 (17.5)	2577 (16.8)	1707 (12.5)	2380 (17.0)
1	2984 (8.4)	1386 (9.0)	790 (5.8)	1267 (9.1)
2	3306 (9.3)	1143 (7.4)	1788 (13.1)	752 (5.4)
3+	6257 (17.6)	1725 (11.2)	3605 (26.5)	783 (5.6)
Missing	16 748 (47.1)	8536 (55.5)	5720 (42.0)	8817 (63.0)
Myocardial infarction	1310 (3.7)	528 (3.4)	425 (3.1)	408 (2.9)
Congestive heart failure	1489 (4.2)	567 (3.7)	395 (2.9)	433 (3.1)
Peripheral vascular disease	632 (1.8)	270 (1.8)	226 (1.7)	205 (1.5)
Cerebrovascular disease	1057 (3.0)	385 (2.5)	308 (2.3)	311 (2.2)
Dementia	608 (1.7)	257 (1.7)	138 (1.0)	194 (1.4)
Chronic obstructive pulmonary disease	1670 (4.7)	673 (4.4)	538 (4.0)	487 (3.5)
Connective tissue/rheumatic disease	193 (0.5)	68 (0.4)	55 (0.4)	54 (0.4)
Peptic ulcer disease	600 (1.7)	233 (1.5)	241 (1.8)	151 (1.1)
Mild liver disease	504 (1.4)	135 (0.9)	147 (1.1)	104 (0.7)
Diabetes without complications	2715 (7.6)	1181 (7.7)	1498 (11.0)	899 (6.4)
Diabetes with complications	1619 (4.6)	647 (4.2)	636 (4.7)	496 (3.5)
Hemiplegia or paraplegia	215 (0.6)	59 (0.4)	55 (0.4)	49 (0.4)
Renal disease	1072 (3.0)	357 (2.3)	277 (2.0)	252 (1.8)
Primary cancer	3159 (8.9)	870 (5.7)	2457 (18.1)	297 (2.1)
Moderate or severe liver disease	318 (0.9)	67 (0.4)	114 (0.8)	52 (0.4)
Metastatic cancer	3251 (9.2)	642 (4.2)	2259 (16.6)	72 (0.5)
HIV/AIDS	16 (<0.1)	≤5 (<0.1)	≤5 (<0.1)	≤5 (<0.1)

Abbreviations: HIV/AIDS, human immunodeficiency virus infection and acquired immune deficiency syndrome.

[a] Data are presented as number (%) or mean ± SD, unless specified otherwise.

[b] *Any diagnosis:* pancreatic cancer was defined by the International Classification of Disease Ninth and Tenth Revision codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using hospital discharge abstracts (any position) or outpatient physician billing and emergency department databases.

Any inpatient diagnosis: pancreatic cancer was defined by ICD-9 or 10 codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using hospital discharge abstracts (any position) only.

Algorithm: pancreatic cancer was defined by ICD-9 or 10 codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using the algorithm defined in Figure 1.

[c] Estimated using the Deyo version of the Charlson Comorbidity Index.[25]

TABLE 2 Properties of the three approaches for identifying pancreatic cancer in Ontario administrative health databases

Approach[a]	Physician/ Hospital Data	Ontario Cancer Registry						
		Recorded	Not Recorded	Total	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)	NPV % (95% CI)
Any diagnosis	Recorded	13 647	21 875	35 522	97.5	99.8	38.4	100.0
	Not recorded	352	12 024 963	12 025 315	(97.2, 97.8)	(99.8, 99.8)	(37.9, 38.9)	(100.0, 100.0)
	Total	13 999	12 046 838	12 060 837				
Any inpatient diagnosis	Recorded	12 122	3245	15 367	86.6	100.0	78.9	100.0
	Not recorded	1877	12 043 593	12 045 470	(86.0, 87.2)	(100.0, 100.0)	(78.2, 79.5)	(100.0, 100.0)
	Total	13 999	12 046 838	12 060 837				
Algorithm	Recorded	10 153	3457	13 610	72.5	100.0	74.6	100.0
	Not recorded	3846	12 043 381	12 047 227	(71.8, 73.3)	(100.0, 100.0)	(73.9, 75.3)	(100.0, 100.0)
	Total	13 999	12 046 838	12 060 837				

Abbreviations: CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value.

[a] *Any diagnosis:* pancreatic cancer was defined by the International Classification of Disease Ninth and Tenth Revision codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using hospital discharge abstracts (any position) or outpatient physician billing and emergency department databases.

Any inpatient diagnosis: pancreatic cancer was defined by ICD-9 or 10 codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using hospital discharge abstracts (any position) only.

Algorithm: pancreatic cancer was defined by ICD-9 or 10 codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using the algorithm defined in Figure 1.

TABLE 3 Differences in timing of pancreatic cancer diagnosis in days between the three approaches using administrative health databases and the Ontario Cancer Registry[a]

Approach[b]	Mean ± SD	Median (IQR)	Minimum	Maximum
Any diagnosis (*N* = 13 647)	43 ± 199	2 (0-21)	0	3662
Any inpatient diagnosis (*N* = 12 122)	49 ± 168	0 (0-28)	0	3149
Algorithm (*N* = 10,153)	55 ± 190	15 (5-37)	0	3662

Abbreviations: IQR, inter-quartile range; SD, standard deviation.

[a] Analyses were restricted to cases identified in both Ontario administrative health data and the OCR. Time was measured as the absolute value of OCR date minus administrative data date in days.

[b] *Any diagnosis:* pancreatic cancer was defined by the International Classification of Disease Ninth and Tenth Revision codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using hospital discharge abstracts (any position) or outpatient physician billing and emergency department databases.

Any inpatient diagnosis: pancreatic cancer was defined by ICD-9 or 10 codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using hospital discharge abstracts (any position) only.

Algorithm: pancreatic cancer was defined by ICD-9 or 10 codes (ICD-9: 157.0-157.9; ICD-10: C25.x) using the algorithm defined in Figure 1.

moderate PPVs (range: 82.0-86.3%).[14] Most recently, Margulis and colleagues found that PPV of 96% for pancreatic cancer patients identified in UK general practitioner records but a sensitivity of only 46% relative to a physician review of patients' medical profile.[17] Our results on the minimal differences in timing of cancer events also are similar to those of other validation studies,[10,12] although neither of these two studies examined the timing of the identification of pancreatic cancer.

Our study has several strengths. First, to our knowledge, ours was the first study to evaluate the validity of different case definitions of incident pancreatic cancer using population-based administrative health data and to compare the approaches to a cancer registry. Ours was also the first such study performed in a Canadian setting; previous studies were primarily conducted in the US using the Surveillance, Epidemiology and End Results-Medicare administrative databases.[7,9-11] Second, the study had a large sample size of 12 000 000 patients that included over 13 500 cases of pancreatic cancer. Finally, we used a reference standard with high levels of case ascertainment (≥90%),[31] so our results should be minimally affected by missing reference standard cases.

Despite these strengths, our study has some limitations worth emphasizing. First, accuracy measures can vary between populations if the patient characteristics, diagnostic coding schemes, or disease prevalences differ.[26,32] Thus, the generalizability of our results to other jurisdictions is unclear, although, in principle, they should be generalizable to other Canadian provinces and territories, which have similar health care systems and data sources. The generalizability of our results to other cancers (eg, breast, colorectal) has also not been established. The specific characteristics of each cancer and its corresponding treatment process determine the validity of any algorithm designed to identify incident cases in claims data. Algorithms should be validated and compared for each type of cancer separately prior to their use in an epidemiologic study. Second, our reference standard was a provincial cancer registry and not original medical records.[26] However, the OCR possesses a wealth of clinical information based upon hospital medical records, cancer clinic records, and pathology reports. It would not have been feasible to validate all pancreatic cancer cases via medical record review. Third, it is possible that pancreatic cancer was not truly incident at the time of diagnosis due to its pathophysiology. To minimize the potential inclusion of prevalent pancreatic cancer, we excluded patients with any history of pancreatic cancer prior to cohort entry. Fourth, the fact that the OCR uses hospital discharge summaries as a data source[33] implies some interdependence of the reference standard with our study databases. This may explain the better performance of the *any inpatient diagnosis* approach relative to the others. Finally, we did not compare severity of cases among those identified using each approach. However, given the expected low survival of patients with pancreatic cancer (8% at 5 years[34]), all pancreatic cancer cases are inherently advanced, and we are less likely to observe differences in the severity of cases than for other cancers (eg, breast, prostate).

In conclusion, among the pancreatic cancer case definitions we studied, the *any inpatient diagnosis* approach was found to be optimal, with the highest PPV, a moderately high sensitivity, and the lowest degree of event date misclassification. Our findings could be useful for future epidemiologic studies of pancreatic cancer.

ACKNOWLEDGEMENTS

This study was funded by CNODES, a collaborating center of the Drug Safety and Effectiveness Network (DSEN), which is funded by the Canadian Institutes of Health Research (Grant Number DSE-146021). We would like to acknowledge the important contributions of the Canadian Network for Observational Drug Effect Studies (CNODES) investigators, collaborators, and assistants. Drs Azoulay and Filion are supported by salary support awards from the *Fonds de recherche du Québec—Santé* (FRQS; Quebec Foundation for Health Research).

This study was supported by the ICES, a non-profit research institute sponsored by the Ontario Ministry of Health and Long-Term Care (MOHLTC). Parts of this material are based on data and information compiled and provided by the Canadian Institute for Health Information (CIHI). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES, the Ontario MOHLTC, or CIHI is intended or should be inferred.

CONFLICT OF INTEREST

Dr Wu is currently an employee at Analysis Group (Montreal, Quebec, Canada). She conducted this work while she was doctoral student at McGill University, prior to her current employment. Mr Secrest is currently an employee of IQVIA (Cambridge, Massachusetts, United States) but conducted this work while an employee of the Centre for Clinical Epidemiology. The other authors have no competing interests to declare.

AUTHORS' CONTRIBUTIONS

L.A., K.B.F., and M.P. designed the study. J.W.W. drafted the study protocol and manuscript, with edits from M.H.S. and K.B.F. L.A., A. H., M.P., F.W., and K.B.F. contributed to the study design. A.H. conducted the data analysis. All authors interpreted the study results, reviewed the manuscript for important intellectual content, and approved of the final version of the manuscript.

ORCID

Laurent Azoulay http://orcid.org/0000-0001-5162-3556
Kristian B. Filion http://orcid.org/0000-0001-6055-0088

REFERENCES

1. Know the facts and statistics. 2011. Available at: http://www.pancreaticcancercanada.ca/site/PageServer?pagename=facingpancreaticcancer_facts. Accessed January 5, 2014.
2. SEER cancer statistics review 1975–2014. 2017. Available at: https://seer.cancer.gov/csr/1975_2014/browse_csr.php?sectionSEL=1&pageSEL=sect_01_table.27.html. Accessed September 24, 2017.
3. Azoulay L, Filion KB, Platt RW, et al. Incretin based drugs and the risk of pancreatic cancer: international multicentre cohort study. *BMJ* (Clinical research ed. 2016;352. i581
4. Cheng W-H, Sadeghi S, Lenz H-J, Hay JW, Barzi A. Comparative effectiveness of FOLFIRINOX (FOL) versus gemcitabine and nab-paclitaxel (GNP) for the first-line treatment of metastatic pancreatic cancer. *J Clin Oncol*. 2016;34(suppl4):306.
5. Maisonneuve P, Lowenfels A. Epidemiology and prospects for prevention of pancreatic cancer. *Pancreatic Cancer* 2nd edition. 2016, pages;1-16.
6. Suissa S, Henry D, Caetano P, et al. Canadian Network for Observational Drug Effect Studies (CNODES). CNODES: the Canadian network for observational drug effect studies. *Open Med*. 2012;6(4):e134-e140.
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 National Cancer Inst*. 2012;104(12):931-940.
8. Deshpande AD, Schootman M, Mayer A. Development of a claims-based algorithm to identify colorectal cancer recurrence. *Ann Epidemiol*. 2015;25(4):297-300.
9. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. *Health Serv Res*. 2004;39(6p1):1733-1749.
10. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between Medicare claims and cancer registry data. *Cancer Causes Control*. 2007;18(5):561-569.
11. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. *Med Care*. 1999;37(5):436-444.
12. Baldi I, Vicari P, Di Cuonzo D, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. *J Clin Epidemiol*. 2008;61(4):373-379.
13. Chawla N, Yabroff KR, Mariotto A, McNeel TS, Schrag D, Warren JL. Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. *Ann Epidemiol*. 2014;24:666-672. 72 e1–2
14. Creighton N, Walton R, Roder D, Aranda S, Currow D. Validation of administrative hospital data for identifying incident pancreatic and periampullary cancer cases: a population-based study using linked cancer registry and administrative hospital data in New South Wales, Australia. *BMJ Open*. 2016;6(7):e011161.
15. Ji J, Sundquist K, Sundquist J, Hemminki K. Comparability of cancer identification among Death Registry, Cancer Registry and Hospital Discharge Registry. *Int J Cancer*. 2012;131(9):2085-2093.
16. Porta M, Costafreda S, Malats N, et al. Validity of the hospital discharge diagnosis in epidemiologic studies of biliopancreatic pathology. PANKRAS II Study Group. *Eur J Epidemiol*. 2000;16(6):533-541.
17. Margulis AV, Fortuny J, Kaye JA, et al. Validation of cancer cases using primary care, cancer registry, and hospitalization data in the United Kingdom. *Epidemiology*. 2018;29(2):308-313.

18. Institutes for Clinical Evaluative Sciences. Data repository. 2016. Available at: https://datadictionary.ices.on.ca/Applications/DataDictionary/Default.aspx. Accessed July 1, 2016.
19. Ontario Ministry of Health and Long-Term Care. *Health Analyst's Toolkit 2012*. In: Division Health System Information Management and Investment Division, edition.2012.
20. King M, Nishri D. *The Ontario Cancer Registry moves to the 21st Century*. In: Cancer Care Ontario, edition.2015.
21. Cancer Care Ontario. Ontario Cancer Registry. Available at: https://www.cancercareontario.ca/en/cancer-care-ontario/programs/data-research/ontario-cancer-registry. Accessed June 6, 2018.
22. Cancer Care Ontario. How we collect cancer registry data. Available at: https://www.cancercareontario.ca/en/dataresearch/accessing-data/technical-information/cancer-registry-datacollection. Accessed June 6, 2018,
23. Holowaty EJ, Norwood TA, Wanigaratne S, Abellan JJ, Beale L. Feasibility and utility of mapping disease risk at the neighbourhood level within a Canadian public health unit: an ecological study. *Int J Health Geogr*. 2010;9(1):21.
24. Walter SD, Birnie SE, Marrett LD, et al. The geographic variation of cancer incidence in Ontario. *Am J Public Health*. 1994;84(3):367-376.
25. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol*. 1994;47(11):1245-1251.
26. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol*. 2012;65(3):343-349 e2.
27. Pancreatic Cancer UK. Policy briefing 2013: the real cost of pancreatic cancer diagnoses via emergency admission. 2013. Available at: https://www.pancreaticcancer.org.uk/media/86662/every-lm_policybriefing-final.pdf. Accessed August 19, 2018.
28. Tobias JS, Hochhauser D. *Cancer and Its Management: John Wiley & Sons*; 2009.
29. Routes of Diagnosis. 2016. Available at: http://www.pancreaticcancer.org.uk/diagnosis. Accessed June 15, 2016.
30. Cancer Care Ontario. Ontario cancer statistics 2016. Available at: https://www.cancercareontario.ca/en/statistical-reports/ontario-cancer-statistics-2016-report-0. Accessed July 1, 2016.
31. Hall S, Schulze K, Groome P, Mackillop W, Holowaty E. Using cancer registry data for survival studies: the example of the Ontario Cancer Registry. *J Clin Epidemiol*. 2006;59(1):67-76.
32. Rothman KJ. *Epidemiology: An Introduction: Oxford University Press*; 2012.
33. Cancer Care Ontario. Ontario Cancer Registry. 2012. Available at: https://www.cancercare.on.ca/ocr/. Accessed January 6, 2015,
34. Canadian Cancer Society. Pancreatic cancer. Available at: http://www.cancer.ca/en/cancer-information/cancer-type/pancreatic/prognosis-and-survival/survival-statistics/?region=on. Accessed June 6, 2018.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wu JW, Azoulay L, Huang A, et al. Identification of incident pancreatic cancer in Ontario administrative health data: A validation study. *Pharmacoepidemiol Drug Saf*. 2018;1–8. https://doi.org/10.1002/pds.4641