Received: 12 November 2021 | Revised: 29 April 2022 | Accepted: 18 May 2022
DOI: 10.1002/hed.27109

ORIGINAL ARTICLE

WILEY

Validated algorithms for identifying timing of second event of oropharyngeal squamous cell carcinoma using real-world data

Shahreen Khair MA[1] | Joseph C. Dort MD, MSc[1,2] | May Lynn Quan MD, MSc[1,2,3] | Winson Y. Cheung MD, MPH[1,2] | Khara M. Sauro PhD[1,2,3] | Steven C. Nakoneshny BSc[4] | Brittany Lynn Popowich BHSc[5] | Ping Liu PhD[1] | Guosong Wu PhD[1,5] | Yuan Xu MD PhD[1,2,3,5]

[1]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

[2]Department of Surgery, Cumming School of Medicine, University of Calgary, North Tower, Foothills Medical Centre, Calgary, Alberta, Canada

[3]Department of Oncology, Cumming School of Medicine, University of Calgary, Tom Baker, Cancer Centre, Calgary, Alberta, Canada

[4]The Ohlson Research Initiative, Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, Alberta, Canada

[5]Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Teaching Research and Wellness (TRW), Calgary, Alberta, Canada

Correspondence
Yuan Xu, Cumming School of Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, Alberta T2N4Z6, Canada.
Email: yuxu@ucalgary.ca

Abstract

Background: Understanding occurrence and timing of second events (recurrence and second primary cancer) is essential for cancer specific survival analysis. However, this information is not readily available in administrative data.

Methods: Alberta Cancer Registry, physician claims, and other administrative data were used. Timing of second event was estimated based on our developed algorithm. For validation, the difference, in days between the algorithm estimated and the chart-reviewed timing of second event. Further, the result of Cox-regression modeling cancer-free survival was compared to chart review data.

Results: Majority (74.3%) of the patients had a difference between the chart-reviewed and algorithm-estimated timing of second event falling within the 0–60 days window. Kaplan–Meier curves generated from the estimated data and chart review data were comparable with a 5-year second-event-free survival rate of 75.4% versus 72.5%.

Conclusion: The algorithm provided an estimated timing of second event similar to that of the chart review.

KEYWORDS

algorithm validation, cancer second event, chart review, oropharyngeal squamous cell carcinoma, real-world data

1 | INTRODUCTION

Oropharyngeal squamous cell carcinoma (OPSCC) is the most common type of cancer to form in the oropharynx and is increasing in incidence in most areas of the world.[1] Most patients with OPSCC are treated with a combination of radiation therapy and chemotherapy, albeit surgery is becoming a more common treatment alternative, especially for patients with early stage, p16+ disease.[2] Second-event-free survival is widely used as key clinical end point is widely used to assess the efficacy of a certain treatment, because it more specifically reflects the treatment effects on cancer outcomes, and also would allow a reduction in the observation duration and cost of the development of new treatments, compared to overall survival.[3]

A second event following curative-intent treatment for primary OPSCC cancer diagnosis can include cancer recurrence and second primary cancer diagnoses. In order to compare treatment outcomes and properly allocate health care resources to inform treatments, decision-makers need data on the timing of second events of head and neck cancer. Traditional chart review is the usual technique for determining second events, but chart review is expensive, time-consuming and not feasible for ascertainment of second events of cancer in large-scale data. Therefore alternative, more cost-effective methods need to be used to allow investigators to explore second events as outcomes in large population-based real-world studies, to better understand OPSCC survival and prognosis.

The processes inherent in delivering health care generate an abundance of data such as physician billing claims, electronic medical records, registry data, and so on. Since such data are generated through the routine delivery of health care it can be considered "real-world" data. Real-world data are therefore easily available and often under-used to understand clinical outcomes and their impact on resource utilization.[4] Many researchers are using real-world data to generate evidence to inform decision making in cancer care. However, effective use of this data requires identification of the timing and nature of second events of head and neck cancer. Despite this need, data on the timing of second events are not explicitly captured in real-world data, resulting in significant difficulties in ascertainment of this variable.[5]

While a number of studies explored the development of algorithms to identify second events of cancer,[6–28] very few explored developing algorithms to establish temporality of second events in the head and neck cancer population. Previous work by Xu et al. developed and validated algorithms to determine second events of OPSCC.[29] Although the previous work provided methods to identify OPSCC patients with second events, it did not determine the timing of second events and thus the length of cancer-free survival, which is essential for survival analysis.[29] In the current paper, we used the previous algorithm to estimate the timing of OPSCC second event and then assessed the validity of the algorithm regarding estimating the timing of second event for cancer-free survival analysis.

2 | METHODS

2.1 | Data and study cohort

All adult (≥18 years) patients with OPSCC curatively treated at the Tom Baker Cancer Centre (Calgary, Canada) (between 1 January 2009 and 31 December 2015) were eligible for inclusion. The Tom Baker Cancer Centre is the sole tertiary cancer center serving Southern Alberta (population 2.4 million). Patients presenting with distant metastases, multiple primary tumors, or not treated with curative intent were excluded from the study. The final observation or the last follow-up date of the patients was 30 September 2017. Patients were censored at the last known follow-up date, if they died before the occurrence of a second event, or had a diagnosis of a second primary non-oropharyngeal tumor. According to a well-established sample-size estimation method,[30] for a 20% second-event incidence, with at least 80% power at the significant level of 0.05 and an estimated 80% sensitivity for the developed algorithm, we need 536 patients. Our cohort met the minimum sample size required. This study used data from Alberta Cancer Registry, National Ambulatory Care Reporting System, Discharge abstract Database, Vital Statistics Database, and provincial physician billing claims. The Alberta Cancer Registry collects data on all patients with a new cancer diagnosis as well as data on cancer deaths.[31] Alberta Cancer Registry provides data on tumor stage, histology, and molecular subtypes. National Ambulatory Care Reporting System contains hospital and community based ambulatory data from day surgeries, outpatient and community-based clinics, and emergency departments.[32] The Discharge abstract Database provides administrative and demographic information on hospital discharges.[33] The vital statistics database collects cause of death and demographic information from provincial and territorial vital registries.[34] Type and date of procedure (e.g., diagnostic imaging, biopsy, and surgery) is recorded in physician claims data.

This study was reviewed and approval by the Health Research Ethics Board of Alberta – Cancer Committee (reference number: HREBA.CC-16-0644). This study only used the secondary data of patients, thus the consent to participate was not required by the ethics review committee.

2.2 | Definition of study variables

The primary outcome variable studied in this paper was second-event free survival. Patients with OPSCC who received an intervention subsequent to their initial treatment were deemed to have a higher probability of a second event than those who did not receive an intervention. Interventions of interest included: chemotherapy or radiation, a diagnostic or surgical procedure, or change in frequency of visits to the cancer center after initial treatment. Table 1 defines all input variables and

TABLE 1 Study variables specifications and ICD codes

Procedure and diagnosis	Codes
Surgery/procedure	
Neck dissection	CCI: 1.GE.91.^ Alberta physician billing codes: 52.31^
Partial/hemi glossectomy	CCI: 1.FJ.87.^ Alberta physician billing codes: 37.1A, 37.1B, 37.2
Lip excision	CCI: 1.YE.87.^ Alberta physician billing codes: 98.6A, 98.6B
Tonsillectomy	CCI: 1.FR.89.^ Alberta physician billing codes: 43.1A
Laryngectomy	CCI: 1.GE.87.^ Alberta physician billing codes: 42.3A, 42.3B, 42.3C
Maxillectomy	CCI: 1.ED.87^, 1.ED.91^ Physician billing codes: 88.4A
Temporal bone subtotal resection	CCI: 1.DR.91^ Physician billing codes: 89.78E
Tracheotomy	CCI: 1.GJ.77^ Physician billing codes: 43.1A
Submandibular gland resection	CCI: 1.FN.87^, 1.FN.89^ Physician billing codes: 38.21A
Laryngoscopy	CCI: 2.GE.70^, 2.GE.71^ Physician billing codes: 42.09B, 42.09D
Primary cancer site (oropharyngeal cancers)	ICD-O: C00.^ - C10.^ and C14.^ ICD-9: 140.^ - 149.^ ICD-10: C00.^ - C10.^ and C14.^
Specialty visits (physician specialty visit only)	Based on the provider classification we identified the type of the physician specialty including the oncologists and head and neck surgeons
Death caused by cancer	ICD-9: 140.^ - 208.^ ICD-10: C00.^ - C97.^

Note: Physician billing codes are derived from the Alberta Schedule of Medical Benefits.

Abbreviations: CCI, Canadian Classification of Health Intervention; ICD-9, International Classification of Disease – ninth edition; ICD-10, International Classification of Disease – tenth edition; ICD-O, International Classification of Disease for Oncology.

their corresponding International Classification of Disease and Canadian Classification of Health Intervention codes.[35] The variables are defined in detail in a 2019 publication by Xu et al. in the Head and Neck Journal and summarized below.[29]

2.2.1 | Subsequent chemotherapy or radiation

The number of chemotherapy episodes which occurred after the primary treatment, within a specific time period (180, 365, and 540 days) were counted. For the purposes of this algorithm, if the number of chemotherapy episodes was equal to or more than three, it was assumed that the patient experienced a second event.[29]

2.2.2 | Diagnostic or surgical procedure

Indicators were created for procedures such as laryngoscope, biopsy, surgery, or radiation therapy following the primary cancer treatment. Patients who encountered an aforementioned intervention were deemed to be at a higher probability of a second event, compared to those patients who did not encounter such interventions.[29]

2.2.3 | Frequent visits to the cancer center

The number of days between two visits to a cancer center was calculated and a new cluster of visits was recorded if they occurred within a prespecified interval of 90, 120, or 180 days. If the number of visits within this cluster was greater than three, four or five encounters, this was considered an indicator of high probability of OPSCC second event.[29]

These variables were used in our previously developed algorithm to determine the status and timing of second event. Each indicator variable in the algorithm can separate the cohort into second-event cases and non-second-event cases, and some of the indicators have a specific date (e.g., the second surgery date) which can be used to estimate the date of the second event for the patient who had the indicator. For indicators without a specific date but contain a timeframe/window (e.g., a new cluster of cancer center visits), we estimated the date of second event as the middle of the timeframe/window, if the patient had the indicator.

2.3 | Statistical analyses

We developed several algorithms with different performance metrics in our previous study; we selected the algorithm with the highest overall accuracy for the current study. Then we applied the algorithm to the entire cohort to identify second event and timing of second event for each patient. Retrospectively collected gold standard chart review data for the entire study cohort were used to validate the timing of second events determined by the algorithm.[29] One investigator (SN), trained by a senior surgeon (JCD), conducted the chart review independently. Ambiguous cases encountered during primary data entry were reviewed by SN and JCD with a consensus being recorded.

Other factors used to develop a Cox-regression model for analyzing second event free survival included patient age at diagnosis, patient sex, year of diagnosis, stage and grade of OPSCC, treatment received, and comorbidities (which were not directly associated with cancer specific survival but could be effect modifiers). Patient age (years), sex (female vs. male), AJCC – 6th pathologic tumor stage (T1-4 and N0-3), tumor grade (1, 2, 3, and unknown), and treatment (surgery yes/no, radiation yes/no, and chemotherapy yes/no) was retrieved from Alberta Cancer Registry. Charlson comorbidities were defined using the established International Classification of Disease algorithm from Discharge Abstract Database, National Ambulatory Care Reporting System, and physician billing claims data and categorized as either 0, 1, or ≥2 comorbidities.[29]

Based on the previously developed algorithm for identifying the status of second event (yes/no) using the classification and regression tree models (CART), the timing of second event was determined by the timing of the indicator event (or variable) that contributed to the identification of the second event.

For the purpose of validating the method developed for estimating the timing of second events, the difference (absolute number in days) between the estimated timing of second event and the chart-reviewed timing of second event was calculated for the second-event patients (determined by chart review), since non-second-event cases had no date of second event for comparison. This difference in timing was then categorized into 0–15, 16–30, 31–60, 61–90, and >90-day windows. In addition, the results of Cox-regression, modeling second event-free survival using the algorithm-estimated second event and timing, was compared with that of the chart review data using the entire cohort. CART and SAS 9.4 software was used to conduct the analyses.[36,37]

Descriptive analyses were performed to show the mean (standard deviation), median (interquartile ranges), and frequency (percent) of the patient and tumor characteristics, treatment, and outcome variables. To assess the difference between the patients in each timing category, characteristics such as tumor, treatment, and outcome variables were assessed using the *T* test/Wilcoxon–Mann–Whitney or chi-square/Fisher exact test for continuous and categorical variables, respectively.

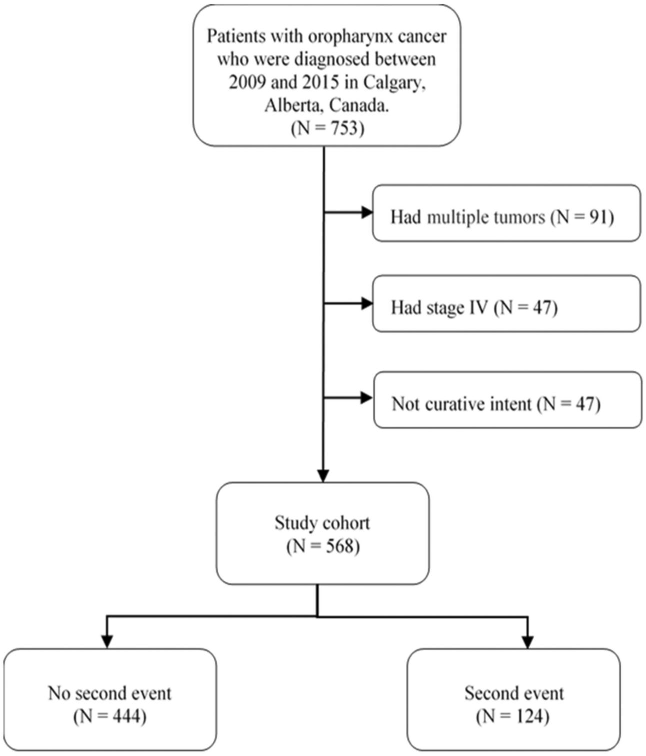
Patients with oropharynx cancer who were diagnosed between 2009 and 2015 in Calgary, Alberta, Canada. (N = 753)
Had multiple tumors (N = 91)
Had stage IV (N = 47)
Not curative intent (N = 47)
Study cohort (N = 568)
No second event (N = 444)
Second event (N = 124)

FIGURE 1 The flowchart of selection of study cohort. Reproduced in exact copy from Xu et al.[29]

3 | RESULTS

A total of 568 (Figure 1) patients with OPSCC that underwent treatment with curative intent were included in the study, with a mean age of 59.4 (standard deviation [SD] 11.5).[29] One hundred and forty-six (25.7%) female patients and 422 (74.3%) male patients were included. Of the entire cohort, 124 (21.8%) patients developed a second event after a median follow-up of 34 (interquartile ratio [IQR] 21–50) months. Most of the patients (96%) were followed for more than 6 months after primary treatment. Based on univariate analysis, the distribution of the variables (Charlson comorbidities, tumor stage, cancer caused death, sex, laterality, pathology tumor grade, radiation, surgery, and status [yes or no] and length of second-event-free survival) were similar in the chart-review determined patient group compared with the algorithm determined patient group. However, the distribution of some variables or characters (including age at diagnosis, tumor stage, chemotherapy, year) were slightly different between the chart reviewed and the algorithm defined groups (Table 2).

The algorithm for identifying second event status (yes vs. no) had 87.9% sensitivity, 84.5% specificity, 61.2% positive predictive value (PPV), 96.2% negative predictive value (NPV), and 85.2% accuracy compared to chart

TABLE 2 The comparison of characteristics between chart review and algorithm determined cohort

Variables	Category	Total based on chart review ($N = 568$)	Chart review determined second event ($N = 124$)	Algorithm estimated second event ($N = 108$)
Age at diagnosis (year)	<50	103 (18.1%)	16 (12.9%)	12 (11.1%)
	50–59	200 (35.2%)	41 (33.1%)	30 (27.8%)
	60–69	158 (27.8%)	39 (31.5%)	35 (32.4%)
	≥70	107 (18.8%)	28 (22.6%)	31 (28.7%)
Sex	Female	146 (25.7%)	28 (22.6%)	30 (27.8%)
	Male	422 (74.3%)	96 (77.4%)	78 (72.2%)
Year of diagnosis	2009–2011	231 (40.7%)	57 (46%)	57 (52.8%)
	2012–2015	337 (59.3%)	67 (54%)	51 (47.2%)
Length of second event free survival based on chart-review (month)	Median (IQR)	28.5 (13.1–53.3)	11 (7.2–19.4)	10.7 (7.1–19.6)
Number of Charlson comorbidities	0	447 (78.7%)	89 (71.8%)	74 (68.5%)
	1	77 (13.6%)	20 (16.1%)	19 (17.6%)
	2+	44 (7.7%)	15 (12.1%)	15 (13.9%)
N classification	N0-1	241 (42.4%)	51 (41.1%)	48 (44.4%)
	N2-3	327 (57.6%)	73 (58.9%)	60 (55.6%)
T classification	T1-2	358 (63%)	62 (50%)	59 (54.6%)
	T3-4	210 (37%)	62 (50%)	49 (45.4%)
Cancer caused death	No	450 (79.2%)	44 (35.5%)	24 (22.2%)
	Yes	118 (20.8%)	80 (64.5%)	84 (77.8%)
Pathology tumor grade	0	561 (98.8%)	123 (99.2%)	106 (98.1%)
	1, 2, 3	7 (1.2%)	1 (0.8%)	2 (1.9%)
Treatment				
Chemotherapy	No	263 (46.3%)	67 (54%)	66 (61.1%)
	Yes	305 (53.7%)	57 (46%)	42 (38.9%)
Radiation	No	153 (26.9%)	33 (26.6%)	32 (29.6%)
	Yes	415 (73.1%)	91 (73.4%)	76 (70.4%)
Surgery	No	304 (53.5%)	48 (38.7%)	29 (26.9%)
	Yes	264 (46.5%)	76 (61.3%)	79 (73.1%)

Abbreviations: CI, confidence interval; T classification, pathological stage of tumor size based on American Joint Committee on Cancer 6th version; N classification, pathological stage of lymph node involvement based on American Joint Committee on Cancer 6th version.

TABLE 3 The validity of the algorithms for identifying second event of oropharyngeal squamous cell carcinoma in the full cohort

Algorithm	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Accuracy, % (95% CI)
High sensitivity	87.9 (82.2–93.6)	84.5 (81.1–87.8)	61.2 (54.1–68.4)	96.2 (94.2–98.1)	85.2 (82.3–88.1)
High PPV	52.4 (43.6–61.2)	99.1 (98.2–100.0)	94.2 (88.7–99.7)	88.2 (85.3–91.0)	88.9 (86.3–91.5)
High accuracy	73.4 (65.6–81.2)	96.2 (94.4–98.0)	84.3 (77.4–91.1)	92.8 (90.5–95.2)	91.2 (88.9–93.5)
Combined method for high sensitivity and high PPV	87.9 (82.2–93.6)	99.1 (98.2–100.0)	96.5 (93.1–99.9)	96.7 (95.1–98.3)	96.7 (95.2–98.1)

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

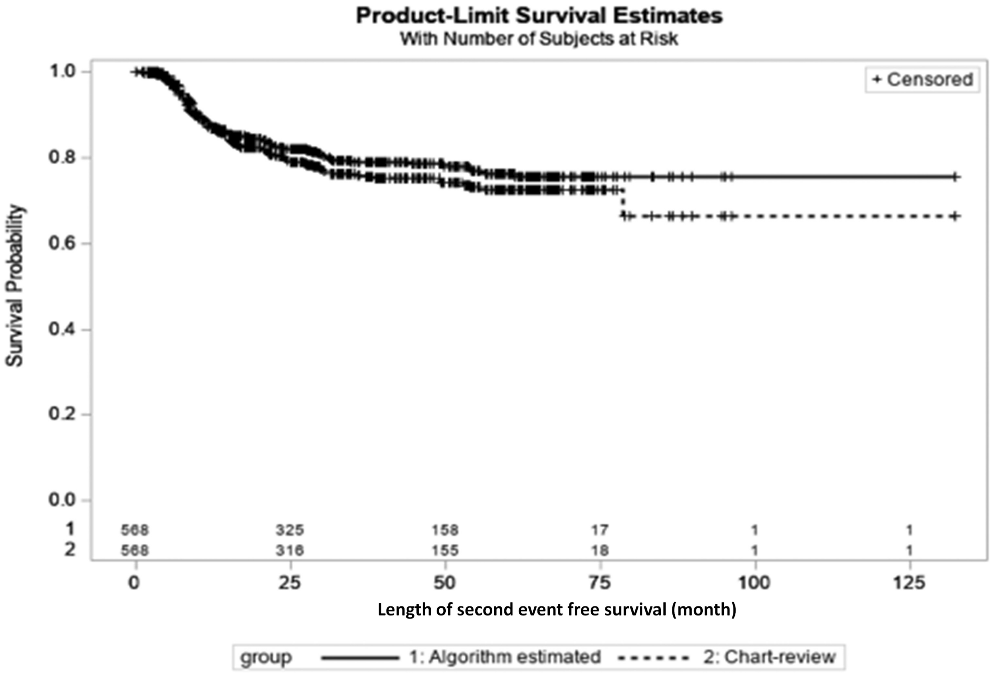

FIGURE 2 Kaplan–Meier curves using chart-review determined and algorithm-estimated data, respectively

TABLE 4 Cox regression models with chart-review data and algorithm-estimated second event data

		Chart-review data		Algorithm-estimated data	
Variable	**Category**	**HR (95% CI)**	***p*-value**	**HR (95% CI)**	***p*-value**
Age (year)	<50	Reference		Reference	
	50–59	1.51 (0.76–2.98)	0.2374	1.44 (0.73–2.86)	0.2938
	60–69	2.23 (1.12–4.41)	0.0216	2.14 (1.08–4.27)	0.0216
	≥70	2.59 (1.27–5.27)	0.0086	2.59 (1.27–5.27)	0.03
CCI	0	Reference		Reference	
	1	1.45 (0.86–2.43)	0.1648	1.39 (0.82–2.34)	0.2162
	≥2	2.25 (1.25–4.08)	0.0071	2.3 (1.27–4.15)	0.0059
Chemotherapy	No	Reference		Reference	
	Yes	0.99 (0.55–1.78)	0.9638	1.05 (0.58–1.91)	0.8722
Sex	Female	Reference		Reference	
	Male	0.99 (0.62–1.58)	0.9791	1 (0.62–1.59)	0.9897
T classification	T1-2	Reference		Reference	
	T3-4	1.31 (0.87–1.96)	0.1989	1.13 (0.75–1.71)	0.5613
N classification	N0-1	Reference		Reference	
	N2-3	1.33 (0.83–2.12)	0.2337	1.34 (0.83–2.16)	0.2275
Radiation*	No	Reference		Reference	
	Yes	1.93 (1.13–3.3)	0.0166	2.14 (1.24–3.68)	0.0062
Surgery*	No	Reference		Reference	
	Yes	5.1 (2.9–8.96)	<0.0001	3.99 (2.23–7.15)	<0.0001
Year	2009–2011	Reference		Reference	
	2012–2015	0.76 (0.51–1.14)	0.1842	0.76 (0.51–1.12)	0.1659

Note: Asterisk (*) indicates that the variable was statistically significant (*p*-value <0.05) based on chart review data.

Abbreviations: CI, confidence interval; CCI, Charlson comorbidity index; HR, hazard ratio; N classification, pathological stage of lymph node involvement based on American Joint Committee on Cancer 6th version; T classification, pathological stage of tumor size based on American Joint Committee on Cancer 6th version.

review (Table 3).[29] The number of patients with a difference between the algorithm-estimated timing of second event and the chart-review determined timing of second event falling within the previously categorized 0–15, 16–30, 31–60, 61–90, and >90-day windows were 43 (34.7%), 25 (20.2%), 24 (19.4%), 10 (8.1%), and 22 (17.7%), respectively. In other words, the majority of patients (74.3%) had a difference falling within the 0–60 day window.

The estimated data and chart review data generated Kaplan–Meier curves (Figure 2) were similar—the 5-year second-event free-survival rate in the chart review data was 72.5% (95% confidence interval [CI]: 70.2–74.7) and 75.4% (95% CI: 73.1–77.7) in algorithm estimated data. Comparing the results of Cox-regression using estimated timing and status of second event with that of chart-review data, there was no significant difference in the hazard ratio (also the *p*-value) of the independent variables (including age, sex, tumor grade, stage, comorbidities, treatments, and year of diagnosis) (Table 4).

4 | DISCUSSION

The methods developed in this study for identifying timing of OPSCC second event achieved high validity, in comparison to chart review. The algorithm established in this study provides a new method to ascertain the timing of OPSCC second event from large-scale, population-based health data to facilitate real-world second-event-free survival analysis.

Understanding and identifying cancer second event is an important endeavor for improving survival and quality of life in patients undergoing a cancer diagnosis.[38] The ability to predict and assess risk for second event has advanced for other types of cancer, such as prostate cancer, but the same progress has not been made for the OPSCC population. Our algorithm fills a gap in the literature and thus a gap in care for patients undergoing treatment for OPSCC by providing a method to detect second events in real-world data and thus understand efficacy of treatment and how to allocate resources to treat and prevent second events in this population. A scan of the literature established a publication by Ricketts et al. as the latest on identifying second events using routinely collected administrative data in this population.[6] It is the only study we found that attempted to identify timing of second event of OPSCC. The methods used to develop the algorithm were, however, limited and sensitivity of the algorithm was low (52.5%). This is probably because their methods mainly relied on diagnostic procedures to indicate that a second event occurred. Patients who did have a second event of OPSCC but did not have evidence of a diagnostic procedure were missed in that study. Our methods differ from theirs as they incorporated the patterns of care trajectory, and hence, may capture second event cases that did not have any diagnostic procedures.

While the current body of literature on developing algorithms to establish second event of cancer in the head and neck population is sparse, lack of comparability extends beyond differences in location of the primary tumor. Studies that focus on algorithms that establish timing of second event of cancer also differ from our research in terms of method, population, region, data, and tumor type.[7–28] While our research is a novel and robust method for identifying second event OPSCC, it can also act as a useful complement to pre-existing research. Algorithms to identify timing of OPSCC second events are less common compared to algorithms designed to do the same for other tumor groups; however, the methods established by this research can be extrapolated to further contribute to other tumor groups.[7–28] In addition, the availability of a robust and accurate algorithm for identifying the timing of OPSCC second event can be used to answer various research questions important to progressing OPSCC detection and treatment.

The model we developed is based on the natural progression and treatment of OPSCC; thus, we believe it can be generalized to most OPSCC patients. Compared to other studies, the algorithm through which the timing of second events are identified in this study is based on commonly found data elements in public administrative data from a universal health system. This eliminates the impact of patient patterns of receiving medical treatment which is otherwise influenced by medical insurance.

The algorithms developed in our previous paper offer different performance measures, thus, provide the option to choose a specific algorithm based on a certain research purpose.[29] A high-sensitivity algorithm can be used for preliminary selection of a second-event cohort for further investigation (e.g., surveillance for second events after primary treatment, identifying a primary cohort of interest to perform detailed chart review on, etc.), while the high-PPV algorithm can be used to identify a cohort of second-event cases for follow-up studies (e.g., assess the effectiveness of a new therapy for second-event of OPSCC). For this study, we selected the high accuracy algorithm (with balanced sensitivity and PPV) to identify timing of second event, in order to effectively demonstrate the validity of the algorithm when used for population-based survival analyses. In other words, we chose this algorithm because we wanted to demonstrate its high accuracy in identifying both true cases and non-cases. This study is timely and important as utilizing this algorithm would result in a savings of resources required to identify and time the second event of cancer.[5]

4.1 | Limitations

Even though this is a population-based study, the sample size is relatively small. Therefore, additional research using a larger cohort with a wide geographic and chronological spectrum may further validate the methods. Although the survival analysis results of the developed algorithm are not significantly different from that of the chart-reviewed data, we were unable to externally validate the algorithm given lack of external data. However, we believe the algorithm is applicable to other data from similar health data systems, given we intentionally only used the common data elements to ensure study generalizability. In addition, our algorithm could be further improved, for example, by including more indicator variables such as p16 status, an important predictor of OPSCC outcome.[39]

4.2 | Conclusion

This study makes important strides in providing a population and data based, real-world algorithm for identifying the timing of second events of OPSCC, which demonstrated comparable results to chart-review in terms of survival analysis. External validation using data from other provinces or different countries is needed to further validate the algorithm.

AUTHOR CONTRIBUTIONS

All authors have contributed substantially to conception and design of the research, acquisition of data, writing and drafting of the manuscript, and approval of the final version of the manuscript and agrees to act as a guarantor of the published version.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Joseph C. Dort https://orcid.org/0000-0002-0858-187X
Khara M. Sauro https://orcid.org/0000-0002-7658-4351
Steven C. Nakoneshny https://orcid.org/0000-0002-8920-0310
Yuan Xu https://orcid.org/0000-0002-7057-1170

REFERENCES

1. Gazawi FM, Lu J, Savin E, et al. Epidemiology and patient distribution of oral cavity and oropharyngeal SCC in Canada. *J Cut Med Surg*. 2020;24(4):340-349.
2. Hobbs AJ, Brockton NT, Matthews TW, et al. Primary treatment for oropharyngeal squamous cell carcinoma in Alberta, Canada: a population-based study. *Head Neck*. 2017;39(11):2187-2199.
3. Michiels S, Le Maître A, Buyse M, et al. Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data. *Lancet Oncol*. 2009;10(4):341-350. doi:10.1016/S1470-2045(09)70023-3
4. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data*. 2019;6:54.
5. Xu Y, Kong S, Cheung WY, et al. Development and validation of case-finding algorithms for recurrence of breast cancer using routinely collected administrative data. *Int J Popul Data Sci*. 2019;19(1):210.
6. Ricketts K, Williams M, Liu Z-W, Gibson A. Automated estimation of disease recurrence in head and neck cancer using routine healthcare data. *Comput Methods Biomech Biomed Eng*. 2014;117(3):412-424.
7. A'mar T, Beatty JD, Fedorenko C, et al. Incorporating breast cancer recurrence events into population-based cancer registries using medical claims: cohort study. *JMIR Cancer*. 2020;6(2):e18143.
8. Rasmussen LA, Jensen H, Virgilsen LF, Jensen JB, Vedsted P. A validated algorithm to identify recurrence of bladder cancer: a register-based study in Denmark. *Clin Epi*. 2018;10:1755-1763.
9. Chubak J, Onega T, Zhu W, Buist DS, Hubbard RA. An electronic health record–based algorithm to ascertain the date of second breast cancer events. *Med Care*. 2017;55(12):e81-e87.
10. Ritzwoller DP, Hassett MJ, Uno H, et al. Development, validation, and dissemination of a breast cancer recurrence detection and timing informatics algorithm. 2017;110(3):273-281.
11. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epi*. 2011;64(8):821-829.
12. Cairncross ZF, Nelson G, Shack L, Metcalfe A. Validation in Alberta of an administrative data algorithm to identify cancer recurrence. *Curr Oncol*. 2020;27(3):343-346.
13. Li Z, Li C, Long Y, Wang X. A system for automatically extracting clinical events with temporal information. *BMC Med Inform Decis Mak*. 2020;20(1):198.
14. Rasmussen LA, Jensen H, Virgilsen LF, et al. Identification of endometrial cancer recurrence—a validated algorithm based on nationwide Danish registries. *Acta Oncol*. 2020;60(4):452-458.
15. Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PLOS one*. 2020;15(10):e0237658.
16. Izci H, Tambuyzer T, Tuand K, et al. A systematic review of estimating breast cancer recurrence at the population level with administrative data. *J Natl Cancer Inst*. 2020;112(10):979-988.

17. Ting W-C, Lu Y-CA, Ho W-C, Cheewakriangkrai C, Chang H-R, Lin C-L. Machine learning in prediction of second primary cancer and recurrence in colorectal cancer. *Int J Med Sci*. 2020;17(3):280-291.
18. Aagaard Rasmussen L, Jensen H, Flytkjær Virgilsen L, Jellesmark Thorsen LB, Vrou Offersen B, Vedsted P. A validated algorithm for register-based identification of patients with recurrence of breast cancer—based on Danish Breast Cancer Group (DBCG) data. *Canc Epi*. 2019;59:129-134.
19. Uno H, Ritzwoller DP, Cronin AM, Carroll NM, Hornbrook MC, Hassett MJ. Determining the time of cancer recurrence using claims or electronic medical record data. *JCO Clin Cancer Inform*. 2018;2:1-10.
20. Mazurowski MA, Saha A, Harowicz MR, Cain EH, Marks JR, Marcom PK. Association of distant recurrence-free survival with algorithmically extracted MRI characteristics in breast cancer. *J Magn Reson Imaging*. 2019;49(7):e231-e240.
21. Zeng Z, Espino S, Roy A, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC Bioinform*. 2018;19(S17):498.
22. Kroenke CH, Chubak J, Johnson L, Castillo A, Weltzien E, Caan BJ. Enhancing breast cancer recurrence algorithms through selective use of medical record data. *J Natl Cancer Inst*. 2015;108(3):djv336.
23. Nicolò C, Périer C, Prague M, et al. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. *JCO Clin Cancer Inform*. 2020;4:259-274.
24. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. *Med Care*. 2017;55(12):e88-e98.
25. Lash TL, Riis AH, Ostenfeld EB, Erichsen R, Vyberg M, Thorlacius-Ussing O. A validated algorithm to ascertain colorectal cancer recurrence using registry resources in Denmark. *Int J Cancer*. 2014;136(9):2210-2215.
26. rHaque R, Shi J, Schottinger JE, et al. A hybrid approach to identify subsequent breast cancer using pathology and automated health information data. *Med Care*. 2015;53(4):380-385.
27. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J Am Med Info Ass*. 2013;20(2):349-355.
28. Carrell DS, Halgrim S, Tran D-T, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epi*. 2014;179(6):749-758.
29. Xu Y, Kong S, Cheung WY, Quan ML, Nakoneshny SC, Dort JC. Developing case-finding algorithms for second events of oropharyngeal cancer using administrative data: a population-based validation study. *Head Neck*. 2019;41(7):2291-2298.
30. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. *J Clin Diagn Res*. 2016;10:YE01-YE06.
31. Alberta Health Services. Alberta Cancer Registry. Alberta Health Services; 2021. https://www.albertahealthservices.ca/cancer/Page17367.aspx
32. National Ambulatory Care Reporting System metadata (NACRS). Canadian Institute for Health Information (CIHI); 2021. https://www.cihi.ca/en/national-ambulatory-care-reporting-system-metadata-nacrs
33. Discharge Abstract Database metadata (DAD). Canadian Institute for Health Information (CIHI); 2021. https://www.cihi.ca/en/discharge-abstract-database-metadata-dad
34. Statistics Canada. Canadian vital statistics—death database (CVSD). Statistics Canada Government of Canada; 2021. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3233
35. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43(11):1130-1139.
36. CART. San Diego, CA: Salford Systems.
37. SAS 9.4. Cary, NC: SAS Institute Inc.; 2014.
38. Achilonu OJ, Fabian J, Bebington B, Elvira S, Eijkemans MJC, Musenge E. Predicting colorectal cancer recurrence and patient survivial using supervicsed machine learning approach: a South African population-based study. *Front Pub Health*. 2021;9:9694306.
39. Ragin CCR, Taioli E. Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: review and meta-analysis. *Int J Cancer*. 2007;121(8):1813-1820. doi:10.1002/ijc.22851

How to cite this article: Khair S, Dort JC, Quan ML, et al. Validated algorithms for identifying timing of second event of oropharyngeal squamous cell carcinoma using real-world data. *Head & Neck*. 2022;44(8):1909-1917. doi:10.1002/hed.27109