Evaluation of algorithms using administrative health and structured electronic medical record data to determine breast and colorectal cancer recurrence in a Canadian province

Kathleen Decker (kdecker@cancercare.mb.ca)
CancerCare Manitoba https://orcid.org/0000-0002-0044-7545
Pascal Lambert
CancerCare Manitoba
Marshall Pitz
CancerCare Manitoba
Harminder Singh
University of Manitoba

Research article

Keywords: breast cancer, colorectal cancer, recurrence, algorithms, validation studies, Canada

Posted Date: October 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-87456/v1

Abstract

Background: Algorithms that use administrative health and electronic medical record (EMR) data to determine cancer recurrence have the potential to replace chart reviews. This study evaluated algorithms to determine breast and colorectal cancer recurrence in a Canadian province with a universal health care system.

Methods: Individuals diagnosed with stage I-III breast or colorectal cancer diagnosed from 2004 to 2012 in Manitoba, Canada were included. Pre-specified and conditional inference tree algorithms using administrative health and structured EMR data were developed. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) correct classification, and scaled Brier scores were measured.

Results: The weighted pre-specified variable algorithm for the breast cancer validation cohort (N=1181, 167 recurrences) demonstrated 81.1% sensitivity, 93.2% specificity, 61.4% PPV, 97.4% NPV, 91.8% correct classification, and scaled Brier score of 0.21. The weighted conditional inference tree algorithm demonstrated 68.5% sensitivity, 97.0% specificity, 75.4% PPV, 95.8% NPV, 93.6% correct classification, and scaled Brier score of 0.39. The weighted pre-specified variable algorithm for the colorectal validation cohort (N=693, 136 recurrences) demonstrated 77.7% sensitivity, 92.8% specificity, 70.7% PPV, 94.9% NPV, 90.1% correct classification, and scaled Brier score of 0.33. The conditional inference tree algorithm demonstrated 62.6% sensitivity, 97.8% specificity, 86.4% PPV, 92.2% NPV, 91.4% correct classification, and scaled Brier score of 0.42.

Conclusions: Algorithms using administrative health and structured EMR data to determine breast and colorectal cancer recurrence had moderate sensitivity and PPV, high specificity, NPV, and correct classification, but low accuracy. Algorithms for determining cancer recurrence must improve before replacing chart reviews.

1. Background

Cancer recurrence is the diagnosis of a second clinical episode of cancer after the first was considered cured. It occurs from residual microscopic disease which was not clinically detectable and is different from cancer progression which is due to the growth of known clinical disease. As novel cancer treatments and screening have been introduced and survival has improved, cancer recurrence has become an important outcome; it often results in additional treatment, is a predictor for subsequent mortality, and can be used to compare treatment effectiveness, measure recurrence-free survival, and plan and prioritize cancer control resources (1).

Since 1956, the Manitoba Cancer Registry (MCR) in the province of Manitoba, Canada has been legislated to collect, classify, and maintain population-based information about cancer cases including diagnosis date, histology, topography, stage, and treatment for the entire provincial population. The MCR has been consistently shown to be of very high quality, completeness, and histological verification (2).

Unfortunately, cancer registries, including the MCR, do not systematically identify recurrent cancers. Therefore, recurrence is determined using manual chart review. Manual chart reviews can provide detailed and reflective information. However, they have several important disadvantages. In order to complete a chart review in a timely manner, multiple abstractors are usually required which may introduce error and bias, especially if data abstraction processes are not clear and inter-rater reliability is low. Manually reviewing charts is also labour intensive, time consuming, and, hence, can be expensive.

An alternative method for identifying recurrence is to use existing structured health care data. Several prior studies have developed and validated algorithms for determining cancer recurrence using structured health data in the United States (US), Europe, and Canada (3–10). Each of these studies has limitations including limited sources of data, narrowly defined populations, different definitions of recurrence, missing validation cohorts, and the use of suboptimal performance measures which do not account for prevalence. Based on the importance of determining recurrence, the inefficiencies of chart reviews, and the limitations of previous recurrence algorithm studies, our goal was to evaluate algorithms to determine recurrence in breast and colorectal cancer cohorts using administrative health and structured electronic medical record (EMR) data in a Canadian province with a universal health care system.

2. Material And Methods

2.1 Setting

The province of Manitoba, located in central Canada, has a population of 1.37 million (as of 2019) (11). Approximately 55% of the population live in the capital city of Winnipeg. Manitoba Health, Seniors and Active Living (MHSAL), the publicly funded provincial health insurance agency, provides comprehensive universal health coverage for hospitalizations, procedures, and physician visits for provincial residents. MHSAL maintains several electronic databases to monitor health care use and reimburse health care providers for services delivered. Since 1984, provincial residents have been assigned a personal health identification number (PHIN) which can be used to link provincial health information databases allowing health care utilization and outcomes to be tracked longitudinally.

2.2 Data sources

The MCR was used to identify individuals diagnosed with breast or colorectal cancer, cancer diagnosis date, age at diagnosis, cancer stage, estrogen receptor/progesterone receptor (ER/PR) and human epidermal growth factor receptor 2 (HER-2) status, date and type of cancer surgery, and date of the first radiation treatment for each course of radiation treatment. The CancerCare Manitoba (CCMB) electronic medical record is the record of clinical cancer interactions, investigations, and treatment, and was used to determine dates and types of systemic anti-cancer medical therapy as well as carcinoembryonic antigen (CEA) and cancer antigen 15 − 3 (Ca 15 − 3) blood test results.

We used three MHSAL administrative databases: the Manitoba Population Registry, the Medical Claims database, and the Drug Programs Information Network (DPIN) database. The Manitoba Population Registry contains demographic, vital status, and migration information and was used to determine the

start and end dates of provincial health coverage. The Medical Claims database is generated by claims filed by health care providers for reimbursement of service and includes services provided, diagnosis, provider, and service date. Medical Claims data were used to determine palliative care consultations. The DPIN database includes all prescriptions dispensed from outpatient pharmacies in Manitoba. DPIN data was used to determine capecitabine, a chemotherapy drug used to treat different cancers including breast and colorectal cancer. Laboratory data were obtained from Shared Health, Manitoba's public sector laboratory, to identify CEA and Ca15-3 blood test results which were not already in the CCMB medical record. The accuracy and completeness of Manitoba Health's administrative data has been previously established (12–14).

2.3 Study population

The study included individuals diagnosed with stage I-III colorectal cancer (International Classification of Diseases, Oncology 3rd edition (ICD-O-3) codes C18.0. C18.2-9, C19, C20, C26.0) or breast cancer (ICD-O-3 codes C50.0-6, C50.8-9). Stage IV cases, which have metastasis at diagnosis, were excluded as these individuals develop progression (i.e., worsening disease) rather than recurrence.

The study population was divided into a training cohort of individuals diagnosed from 2004 to 2007 and a validation cohort of individuals diagnosed from 2008 to 2012. Breast and colorectal cancers were analyzed separately. The breast cancer training cohort included cancers that were either ER negative, PR negative, or HER-2 positive because these cancers have a higher recurrence rate and therefore decreased the number of cases needed to review (15). The colorectal cancer training cohort focused on stage II and III because they are expected to have higher rates of recurrence compared to stage I cancers (16). The breast and colorectal cancer validation cohorts included individuals diagnosed with stage I-III cancers. However, the breast cancer cohort was oversampled with ER negative, PR negative, and HER-2 cases and the colorectal cancer cohort was oversampled with higher stages to ensure that enough recurrences were identified. The validation cohort included individuals diagnosed in later years to provide external validation, which is a more rigorous validation method than internal or apparent validation (17–19).

2.4 Study variables

Study variables are summarized in Table 1. Surgery and radiation treatment data were linked by a tumour ID which identifies the treatment associated with a specific tumour. Therefore, if an individual had more than one cancer diagnosis, the treatment data could be linked to the appropriate cancer. The remaining variables could not be linked to a specific tumour. To increase accuracy in classifying the remaining variables, conditions were added. The receipt of chemotherapy beyond 12 months of diagnosis was considered due to recurrence unless it occurred after a second primary treated with surgery. Although chemotherapy treatment could not be linked to a specific tumour, the treatment site was identified which increased accuracy of correct association. A palliative care consult was considered due to recurrence if it was provided by an oncologist and was linked to a breast cancer (breast cancer cases only), colorectal cancer (colorectal cancer cases only), lung cancer, liver cancer, or undetermined metastases beyond 6 months after diagnosis. Elevated blood markers more than 12 months after diagnosis (CEA > 10; Ca 15 –

3 > 50) were considered due to recurrence unless they occurred within three months of another primary cancer diagnosis.

Table 1
Variables included in the study

Variable	Data source	Time frame after diagnosis
Breast Cancer		
Surgery (mastectomy, lumpectomy, axillary lymph node dissection)	Manitoba Cancer Registry	12 months
Surgery (non-breast site)	Manitoba Cancer Registry	6 months
Chemotherapy	CancerCare Manitoba EMR[3] Drug Programs Information Network	12 months[1]
Radiation therapy	Manitoba Cancer Registry	12 months
Palliative care consultation from oncologist	Medical Claims	6 months
CEA > 10	CancerCare Manitoba EMR Shared Health	12 months[2]
Ca 15 − 3 > 50	CancerCare Manitoba EMR Shared Health	12 months[2]
Colorectal cancer		
Surgery (bypass or resection)	Manitoba Cancer Registry	12 months
Surgery (liver or lung resection)	Manitoba Cancer Registry	6 months
Chemotherapy	CancerCare Manitoba EMR Drug Programs Information Network	12 months[1]
Palliative care consultation from oncologist	Medical Claims	6 months

1. Excluding individuals who had chemotherapy after a second primary cancer diagnosis treated with surgery.

2. Excluding individuals who had elevated blood markers within 3 months of another primary cancer diagnosis.

3. Electronic medical record

Variable	Data source	Time frame after diagnosis
CEA > 10	CancerCare Manitoba EMR Shared Health	12 months[2]
1. Excluding individuals who had chemotherapy after a second primary cancer diagnosis treated with surgery.		
2. Excluding individuals who had elevated blood markers within 3 months of another primary cancer diagnosis.		
3. Electronic medical record		

2.5 Algorithm development and validation

A chart review was first conducted by trained research assistants to identify cancer recurrence. Cancer recurrence included loco-regional (reappearance of cancer in the same region of the body or the lymph nodes) and distant (reappearance of cancer in another part of the body) recurrence. A duplicate chart review by a research assistant who did not conduct the initial chart review was conducted for a fraction of the cohort to evaluate inter-rater reliability. The algorithms were then developed by analyzing the same cohorts using two approaches: pre-specified variables and conditional inference trees. The pre-specified variable approached used variables and clinically meaningful cut offs determined prior to the start of the study. Variables and cut-offs were selected with information from previous studies and local cancer experts. The conditional inference tree approach (an automated machine learning technique) used the same variables as the pre-defined algorithm. However, trees were created based on the association between each covariate and the outcome of interest (i.e., recurrence). The *ctree* function with the party R package was used (20). Validation cohorts were used to determine if the algorithms developed were generalizable to cancer cohorts independent of those analyzed as part of the algorithm development.

2.6 Performance metrics

Sensitivity (the percentage of individuals who had a recurrence that were correctly identified), specificity (the percentage of individuals who did not have a recurrence that were correctly identified), positive predictive value (PPV) (percentage of individuals predicted to have recurrence that truly have recurrence), negative predictive value (NPV) (percentage of individuals predicted to not have recurrence that truly do not have recurrence), correct classification (the percentage of individuals who were correctly classified as having a recurrence or not having a recurrence), and scaled Brier scores were calculated to determine algorithm accuracy. Brier scores are the average of squared differences between predicted values and outcome values. The Brier score was then scaled to the proportion of events (p) in the cohort (1 – (Brier score /(mean)(p)*(1-mean(p)))) where a value of 1 is perfect prediction, a value of 0 is chance, and a negative value is worse than chance (21, 22). Therefore, unlike measures like sensitivity, specificity, and correct classification, the scaled Brier score considers the rate of events. Measures were unweighted for

both training and validation cohorts. Weighted measures were also calculated for the validation cohort to account for oversampling.

3. Results

3.1 Breast cancer

The breast cancer training cohort included 933 cases with 186 recurrences and the validation cohort included 1181 cases with 167 recurrences (Table 2). The mean age at diagnosis was 60.3 years (standard deviation (SD) 14.1) in the training cohort and 62.5 years (SD 13.9) in the validation cohort. In the training cohort, 36.2% were diagnosed at stage I, 45.7% at stage II, and 18.1% at stage III. The stage distribution was similar in the validation cohort. Two-hundred and seven charts were reviewed by a second research assistant. The Kappa statistic for recurrence status was 0.81.

Table 2
Characteristics of individuals diagnosed with breast or colorectal cancer in the training and validation cohorts

	Breast cancer, N, (%)		**Colorectal cancer, N, (%)**	
	Training cohort (N = 933)	Validation cohort (N = 1811)	Training cohort (N = 620)	Validation cohort (N = 693)
Age at diagnosis (mean, (SD))	60.3 (14.1)	62.5 (13.9)	69.4 (12.6)	67.9 (12.3)
Stage				
I	338 (36.2)	487 (41.2)	0 (0)	138 (19.9)
II	426 (45.7)	485 (41.1)	285 (46.0)	236 (34.1)
III	169 (18.1)	209 (17.7)	335 (54.0)	319 (46.0)
Recurrences	186 (19.9)	167 (9.2)	126 (20.3)	136 (19.6)

Table 3 shows the performance metrics for determining breast cancer recurrence for the training and validation cohorts, unweighted and weighted, using pre-specified variable and conditional tree algorithms. In the validation cohort, the pre-specified variable algorithm demonstrated the following weighted results: 81.1% sensitivity, 93.2% specificity, 61.4% PPV, 97.4% NPV, and 91.8% correct classification. The weighted scaled Brier score was 0.21 which demonstrates low accuracy. The conditional inference tree algorithm for the training cohort is shown in Fig. 1. The percentage of individuals who were classified as having had a recurrence are shown for each node in the tree. For example, 70% of individuals who did not have chemotherapy but had radiation therapy greater than two months after diagnosis were classified as having had a recurrence (node 3, N = 53). Overall, the conditional inference tree algorithm demonstrated

the following weighted results: 68.5% sensitivity, 97.0% specificity, 75.4% PPV, 95.8% NPV, and 93.6% correct classification. The weighted scaled Brier score was 0.39.

Table 3
Performance metrics for determining breast cancer recurrence using pre-specified variable algorithm and conditional tree algorithms

	Pre-defined variable algorithm, %			**Conditional tree algorithm, %**		
	Training cohort (N = 933)	Validation cohort (N = 1811)		Training cohort (N = 933)	Validation cohort (N = 1811)	
		Unweighted	Weighted		Unweighted	Weighted
Sensitivity	83.3	83.2	81.1	78.0	73.7	68.5
Specificity	89.7	92.5	93.2	95.7	96.4	97.0
PPV[1]	66.8	64.7	61.4	78.0	73.7	75.4
NPV[2]	96.0	97.1	97.4	94.6	96.0	95.8
Correct classification	88.4	91.1	91.8	92.2	93.1	93.6
Scaled Brier	0.27	0.27	0.21	0.51	0.44	0.39
1. Positive predictive value						
2. Negative predictive value						

3.2 Colorectal cancer

The colorectal cancer training cohort included 620 cases with 126 recurrences and the colorectal cancer validation cohort included 693 cases with 136 recurrences (Table 2). The mean age at diagnosis was 69.4 years (SD 12.6) in the training cohort and 67.9 years (SD 12.3) in the validation cohort. In the training cohort, no individuals were included who were diagnosed at stage I; 46.0% were diagnosed at stage II and 54.0% were diagnosed at stage II. In the validation cohort, 19.9%, 34.1%, and 46.0% were diagnosed at stages I, II, and III, respectively. One-hundred and twenty-eight charts were reviewed by a second research assistant. The Kappa statistic for recurrence status was 0.73.

Table 4 shows the performance metrics for determining colorectal cancer recurrence for the training and validation cohorts, both unweighted and weighted, using the pre-specified variable and conditional tree algorithms. In the validation cohort, the pre-specified variable algorithm demonstrated the following weighted results: 77.7% sensitivity, 92.8% specificity, 70.7% PPV, 94.9% NPV, and 90.1% correct classification. The weighted scaled Brier score was 0.33. The conditional inference tree algorithm

developed demonstrated the following weighted results: 62.6% sensitivity, 97.8% specificity, 86.4% PPV, 92.2% NPV, and 91.4% correct classification (Fig. 2). The weighted scaled Brier score was 0.42.

Table 4
Performance metrics for determining colorectal cancer recurrence using pre-defined variable algorithms and conditional inference tree algorithms

	Pre-defined variable algorithm, %			**Conditional tree algorithm, %**		
	Training cohort (N = 933)	Validation cohort (N = 1811)		Training cohort (N = 620)	Validation cohort (N = 693)	
		Unweighted	Weighted		Unweighted	Weighted
Sensitivity	88.1	79.4	77.7	71.4	64.7	62.6
Specificity	89.9	92.5	92.8	96.2	96.9	97.8
PPV[1]	68.9	72.0	70.7	82.6	83.8	86.4
NPV[2]	96.7	94.8	94.9	93.0	91.4	92.2
Correct classification	89.5	89.9	90.1	91.1	90.6	91.4
Scaled Brier	0.35	0.36	0.33	0.45	0.41	0.42
1. Positive predictive value						
2. Negative predictive value						

4 Discussion

4.1 Main findings

We found that algorithms using administrative health and structured EMR data to determine breast and colorectal cancer recurrence had high to moderate sensitivity and PPV, high specificity, NPV, and correct classification but low accuracy after adjusting for the prevalence of the outcome in the cohort. As expected, training cohort results were higher than validation cohort results because the algorithms were optimized on the training cohorts. We chose to include breast and colorectal cancers as these sites have relatively high survival rates, are the second and third most commonly diagnosed cancers in Manitoba (which makes chart reviews even more costly and time-consuming), and are historically more likely than aggressive cancers with poorer survival to have recurrences that can be effectively treated. Whether or not these algorithms can replace chart reviews for determining cancer recurrence necessitates weighing the costs required to conduct a chart review with the benefit of quickly applying an algorithm with less than optimal accuracy.

4.2 Comparison with other studies

Prior studies that evaluated cancer recurrence algorithms using structured data found moderate to high sensitivities and specificities but have several important limitations. Lamont et al. (2006) used Medicare claims data to measure disease-free survival in individuals ≥ 65 years of age diagnosed with breast cancer (N = 52, 15 recurrences) (3). Algorithm sensitivity and specificity were 83% and 97%, respectively. A more recent study (2016) developed a medical claims-based algorithm to identify ovarian cancer recurrence (N = 94, 32 recurrences) (4). Sensitivity was 100% and specificity was 89% but only a training cohort was assessed. Because a validation cohort was not included, the generalizability of the study was not evaluated. Both studies had small cohort sizes. Chubak et al. (2012) developed algorithms to determine recurrence among women diagnosed with stage I or II breast cancer (n = 3,152, 407 recurrences) (5). Sensitivity (89% and 96%) and specificity (99% and 95%) were higher than our study. However, they did not distinguish between cancer recurrence and a second primary (i.e., a new primary cancer unrelated to the prior cancer). This distinction is important in order to use the algorithms to evaluate outcomes such as the effectiveness of treatments in preventing a cancer recurrence. We attempted to distinguish recurrence from a second primary in our chart reviews, although this was difficult in some cases. Another large US study (2014) evaluated recurrence algorithms for lung, colorectal, breast, and prostate cancer (n = 6,227, 736 recurrences) (7). Sensitivity ranged from 75–85%. In 2017, the study was extended to include additional data; the AUROC score was > 0.92. (8). Rasmussen et al. (2019) used data national data in Denmark to identify breast cancer recurrence (n = 471, 149 recurrences) (6). Sensitivity was 97.3%, specificity was 97.2%, and PPV was 94.4%. These studies also did not distinguish between recurrence and second primaries and results were optimized on the training cohort which would have produced overly optimistic results.

To our knowledge, only two Canadian studies has developed and validated algorithms for identifying cancer recurrence using administrative health data. Xu et al. (2019) developed algorithms to identify breast cancer recurrence among women ≤ 40 years of age or those who received neoadjuvant chemotherapy in Alberta (N = 598, 121 recurrences) (9). They found higher measures of sensitivity (94.2%) and PPV (93.4%) and similar measures of specificity (98.3%) and NPV (98.5%) compared to our study. They also did not distinguish between a recurrence and a second breast cancer primary and excluded patients with second primary non-breast tumours. This may have introduced bias and reduces the generalizability of the algorithms. Cairncross et al. (2020) randomly selected 200 women (26 recurrences) diagnosed with cancer and who had ever had a pregnancy between 2003 and 2012 (10). Sensitivity was higher than our study (80.8%), specificity and PPV were lower (81.0% and 38.9%), and NPV was similar (96.6%). However, the data used to determine recurrence in this study was incomplete (e.g., hospitalizations were not included) and the study was limited to women of reproductive age.

Importantly, none of the prior studies from the US or Canada used metrics that are optimal to measure algorithm performance such as the scaled Brier score. Sensitivity and specificity are useful because they provide context about how an algorithm can be improved by identifying areas of weakness. For example, some recurrences in our study were missed because the individual did not receive treatment, which

decreased sensitivity. In addition, chemotherapy for a second primary was often found among false positives, which decreased specificity. However, sensitivity and specificity ignore the rate of events in a cohort which make assessing the overall performance of an algorithm challenging. For example, a specificity or correct classification of 95% will have a high rate of false positives if the rate of event is low (e.g., 1%) but would be substantially better with a higher rate of events (e.g., 50%). The scaled Brier score, which is a summary measure that accounts for the rate of events in a cohort, does not have this limitation. Moreover, if a proposed algorithm is expected to replace a chart review, metrics of accuracy should also indicate the amount of measurement error involved. The scaled Brier score, which has a similar interpretation to the R^2, indicates random association with a value of 0 and perfect prediction with a value of 1. This provides more informative output to describe accuracy than measures that only use subsections of the cohort (e.g., sensitivity and specificity).

Other methods, such as those that use natural language processing (NLP) to capture recurrence from unstructured EMR data, have been used to determine breast cancer recurrence with sensitivities from that range from 83–92% (23–26). These results are not very different that those that used structured administrative data and therefore, may also not be accurate enough at this time to replace a chart review. Another option is to use recurrence algorithms as a screening tool to reduce the number of charts that need to be manually reviewed. However, more research is needed to create algorithms with higher sensitivities to evaluate this possibility.

4.3 Strengths and limitations

We used data from previously validated, high-quality, complete, population-based administrative health databases (12, 13, 27, 28). However, our gold standard was a chart review which is subject to human error. The inter-rater reliability was strong for breast cancer and moderate for colorectal cancer (29). Therefore, there was some disagreement among the chart reviewers about what constitutes recurrence. When investigating samples of false positives and negatives in the training cohort, some misclassifications of recurrence status had occurred. This was often due to the difficulty in distinguishing recurrence from a second primary, which is expected because it is sometimes challenging for physicians to definitively make this determination. We also found that additional chemotherapy may have been due to a second cancer primary and not cancer recurrence leading to false positive cases. Like some prior studies, our definition of recurrence was not time dependent. We chose to not include this because this would only lead to poorer results.

5. Conclusions

Our algorithms that used structured administrative health and EMR data to determine recurrence in breast and colorectal cancer cohorts had moderate sensitivity and PPV, high specificity, NPV, and correct classification but low overall accuracy. These results suggest that more accurate algorithms are required to replace chart reviews.

List Of Abbreviations

CCMB CancerCare Manitoba

CEA Carcinoembryonic antigen

DPIN Drug Program Information Network

EMR Electronic medical record

ER/PR Estrogen receptor/progesterone receptor

HER-2 human epidermal growth factor receptor 2

ICD-0-3 International Classification of Diseases, Oncology 3rd edition

MCR Manitoba Cancer Registry

MHSAL Manitoba Health Seniors and Active Living

NPV Negative predictive value

PHIN Personal Health Information Number

PPV Positive predictive value

SD Standard deviation

US United States

Declarations

Consent for publication

All authors consent to the publication of this manuscript.

Availability of data and materials

The data that support the findings of this study are not publicly available to ensure and maintain the privacy and confidentiality of individuals' health information. Requests for may be made to the appropriate data stewards (Manitoba Health, Seniors and Active Living's Health Information Privacy Committee and CancerCare Manitoba's Research and Resource Impact Committee).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the CancerCare Manitoba Foundation. The funders had no role in this study.

Authors' contributions

PL contributed conceptualization, methodology, formal analysis, investigation, data curation, writing – review & editing, visualization, supervision, funding acquisition. MP contributed conceptualization, methodology, resources, writing – review & editing, funding acquisition. HS contributed conceptualization, methodology, resources, writing – review & editing, funding acquisition. KD contributed conceptualization, methodology, writing – original draft, writing – review & editing, visualization, supervision, funding acquisition.

Acknowledgements

We gratefully thank from the CancerCare Manitoba Foundation for their support and CancerCare Manitoba, Manitoba Health, Seniors and Active Living, and Shared Health for the provision of data. We also thank Zack Hall, LoriAnn Love, Coreen Hildebrand, and Megan Noonan for conducting the chart reviews and Elizabeth Harland for study coordination.

References

1. Yu X. Epidemiology of Cancer Recurrence, Second Primary Cancer, and Comorbidity Among Cancer Survivors. Feuerstein M, Ganz P, editors. New York, NY: Springer 2011.
2. North American Association of Central Cancer Registires. APPENDIX C - Data Quality Indicators by Year and Registry. In: Hotes Ellison J, Wu XC, McLaughlin C, Lake A, Firth R, et al., editors. Cancer In North America: 1999-2003 Volume One: Incidence: North American Association of Cancer Registries Inc.; 2006. p. II-325.
3. Lamont EB, Hernon 2nd JE, Weeks JC, Henderson C, Earle CR, Schilsky RL, et al. Measuring disease-free survival and cancer relapse using medicare claims from CALGB breast cancer trial participants (Companion to 9344). J Natl Cancer Inst. 2006;98(18).
4. Livaudais-Toman J, Franco R, Prasad-Hayes M, Howell EA, Wisnivesky J, Bickell NA. A validation of administrative claims data to measure ovarian cancer recurrence and secondary debluking surgery. EGEMS. 2016;4(1).
5. Chubak J, Yu O, Pocobelli G, Lamerato L, Webster J, Prout MN, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive cancer. J Natl Cancer Inst. 2012;104(12):931-40.
6. Rasmussen LA, Jensen H, Flytkjaer Virgilsen L, Beck Jellesmark Thorsen L, 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. Cancer Epidemiology. 2019;59:129-34.
7. Hassett MJ, Ritzwoller DP, Taback N, Carroll N, Cronin AM, Ting GV, et al. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using 2 large

contemporary cohorts. Med Care. 2014;52(10):e65-73.

8. Hassett MJ, Uno H, Cronin AM, Hornbrook MC, Ritzwoller DP. 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.
9. Xu Y, Kong S, Cheung WY, Bouchard-Fortier A, Dort JC, Quan H, et al. Development and validation of case-finding algorithms for recurrence of breast cancer using routinely collected administrative data. BMC Cancer. 2019;19(210).
10. 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):e343-e6.
11. Manitoba Health, Seniors and Active Living. Population Report, June 1, 2019. Winnipeg, MB; 2019.
12. Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease. Comparing administrative data and self-reports. Med Care. 1997;35(9):932-47.
13. Roos LL, Mustard CA, Nicol JP, McLarran DF, Malenka DJ, et al. Registries and administrative data: organization and accuracy. Med Care. 1993;31(3):201-12.
14. Roos LL, Walld R, Uhanova J, Bond R. Physician visits, hospitalizations, and socioeconomic status: ambulatory case sensitive condintions in a Canadian setting. Health Serv Res. 2005;40(4):1167-85.
15. Arvold ND, Taghaian AG, Niemierko A, Abi Raad RF, Srreedhara M, Nguyen PL, et al. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. J Clin Oncol. 2011;29(29):3885-91.
16. Siegel R, Miller K, Fedewa S, Ahnen D, Meester R, Barzi A, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin. 2017;67:177-93.
17. Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc A. 1995;158(3):419-66.
18. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. J Am Stat Assoc. 1983;78(382):316-31.
19. Harrell FE, Lee K, Mark D. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing erros. Stat Med. 1996;15(4):361-87.
20. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat. 2006;15(3):651-74.
21. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1).
22. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and udating. New York, NY: Springer-Verlag; 2009.
23. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. JCO Cllin Cancer Inform. 2019:1-12.

24. Carrell DS, Halgrim S, Tran D-T, Buist DS, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol. 2013;179(6):749-58.
25. Ling AY, Kurian AW, Caswell-Jin JL, Sledge Jr GW, Shah NH, Tamang SR. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open. 2019;2(4):528-37.
26. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018;19:498-508.
27. Roos LL, Traverse D, Turner D. Delivering prevention: the role of public programs in delivering care to high-risk populations. Medical Care. 1999;37(6):JS264-JS78.
28. McPhee SJ, Nguyen TT, Shema SJ, Nguyen B, Somkin C, Vo P, et al. Validation of recall of breast and cervical cancer screening by women in an ethcinally diverse population. Preventive Medicine. 2002;35:463-73.
29. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276-82.

Figures

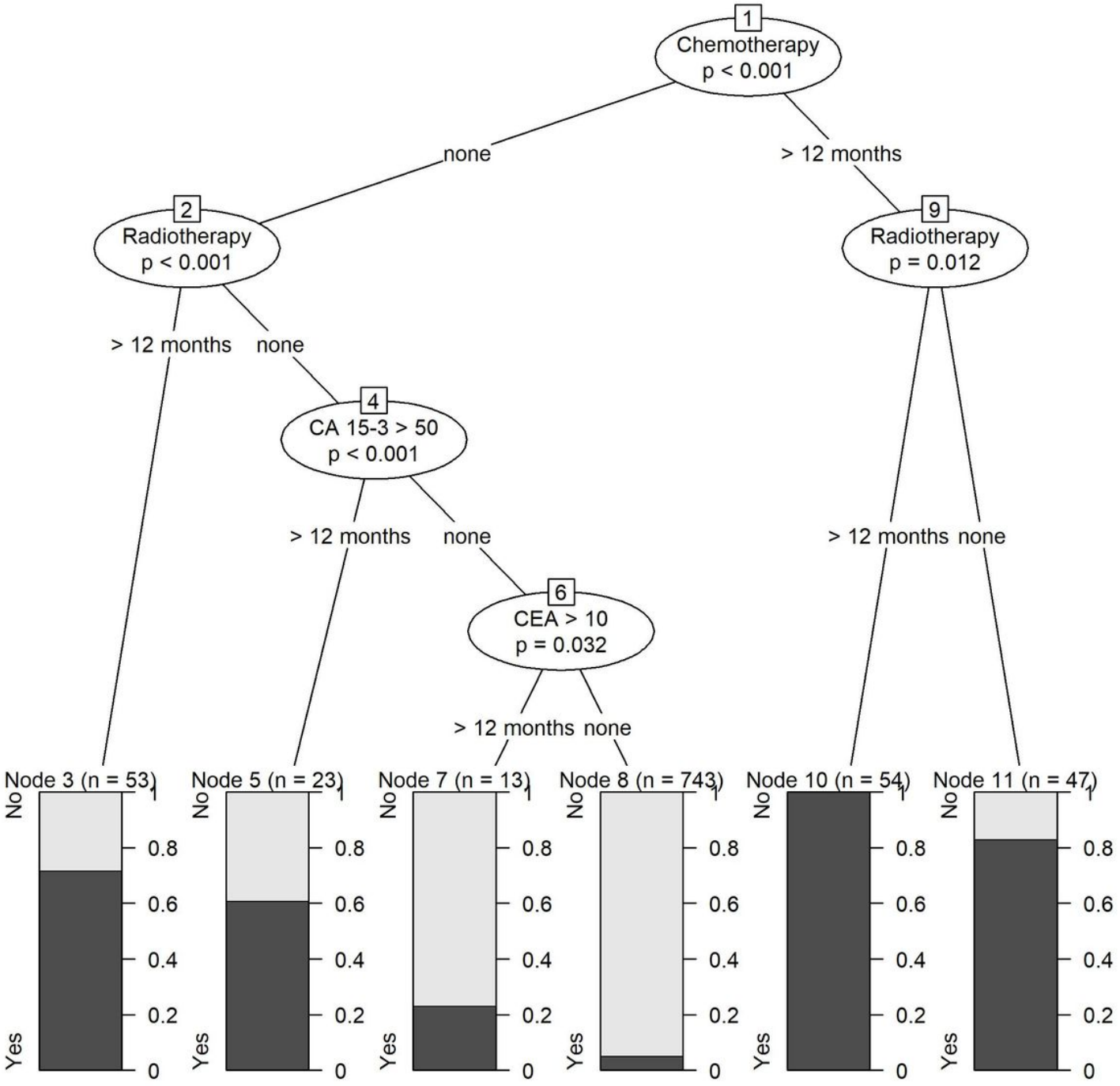

Figure 1

Conditional inference tree for breast cancer recurrence

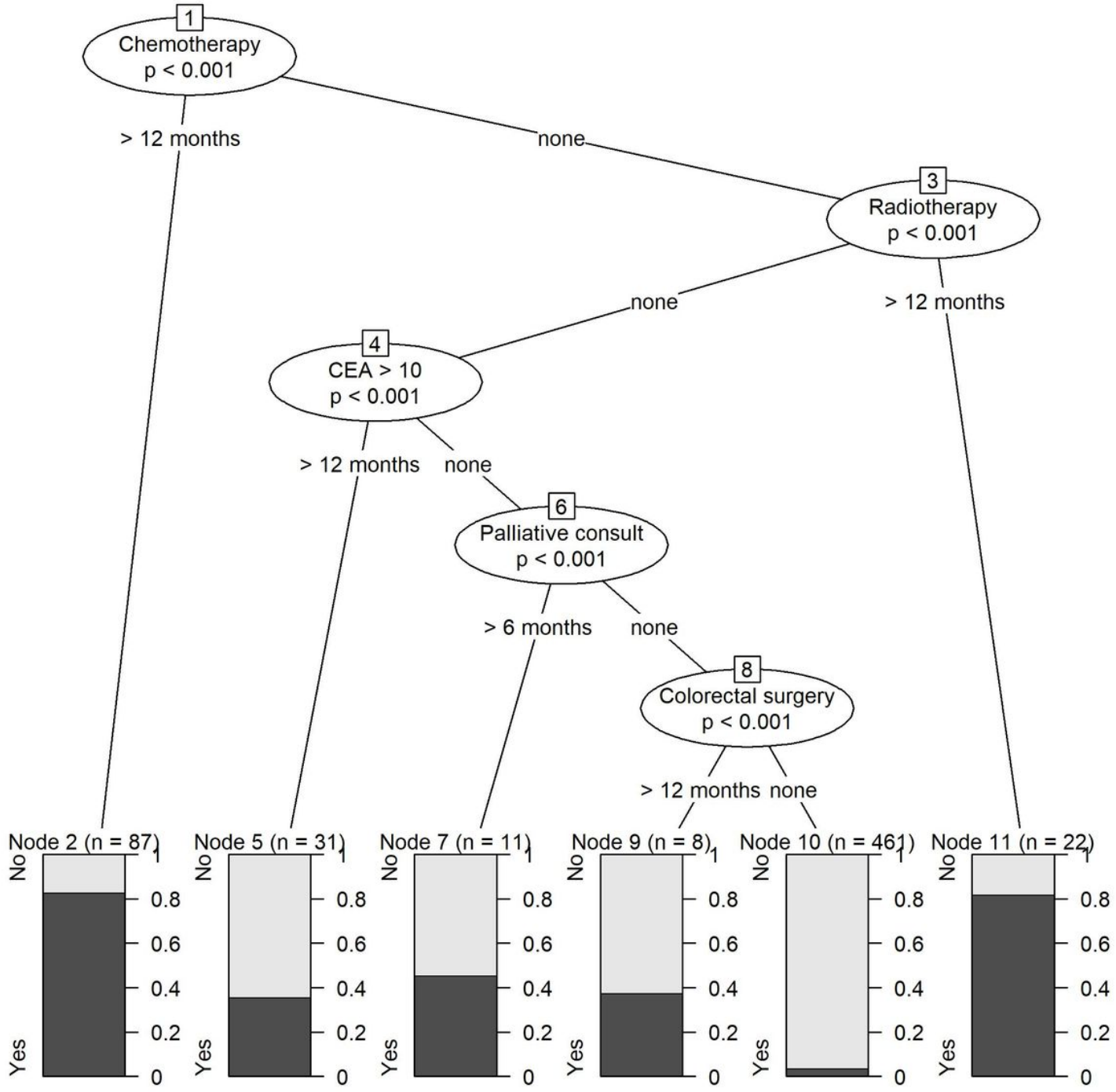

Figure 2

Conditional inference tree for colorectal cancer recurrence