Winter 2025 | Volume 52 | Number 4

Journal of Registry Management

Published by the National Cancer Registrars Association
Founded in 1975 as *The Abstract*

ISSN: 1945-6123

The Right Job for the Ideal Candidate

JOB BANK IS NCRA'S RESOURCE CONNECTING EMPLOYERS TO CANCER REGISTRARS.

OPPORTUNITY AWAITS!

Visit Job Bank today at
www.ncra-usa.org/jobs.

Journal of Registry Management

is published quarterly by the
National Cancer Registrars Association
1330 Braddock Place, #520
Alexandria, VA 22314
(703) 299-6640
(703) 299-6620 FAX

Address any correspondence to:
National Cancer Registrars Association
1330 Braddock Place, #520
Alexandria, VA 22314

Address all editorial correspondence to:
Nadine Walker, MS, ODS-C
Email: JRMEditor@ncra-usa.org
Telephone: (703) 299-6640 ext. 327

Letters to the Editor
Letters to the Editor must be signed and include address and telephone number; none will be published anonymously. Letters subject to editing.

Editorial Policy
Opinions expressed in articles appearing in *Journal of Registry Management* are those of the authors and not necessarily those of the National Cancer Registrars Association, or the editor of *Journal of Registry Management*.

Accessing the *JRM*
Access to the *JRM* is available via NCRA's website at the *JRM* Library (ncra-usa.org). Full-text articles are also available on the National Library of Medicine's PubMed Central website.

Advertising Sales
Advertising is accepted with the understanding that products and services are professionally related and meet the ethical standards of public health practice. Acceptance by the *Journal* does not indicate or imply endorsement by the *Journal* or NCRA. Advertising sales should be directed to Marketing Coordinator, NCRA, 1330 Braddock Place, #520, Alexandria, VA 22314, (703) 299-6640, (703) 299-6620 FAX.

National Cancer Registrars Association

Journal of Registry Management

Winter 2025 • Volume 52 • Number 4

Contents

—continued on page 107

Indexing

The *Journal of Registry Management* is indexed in the National Library of Medicine's MEDLINE database. Citations from the articles indexed, the indexing terms (key words), and the full article are included and searchable using PubMed.

For your convenience, the *Journal of Registry Management* is indexed in the 4th issue of each year and on the Web (under "Resources" at http://www.ncra-usa.org/jrm). The 4th issue indexes all articles for that particular year. The Web index is a cumulative index of all *JRM* articles ever published.

Journal of Registry Management

is published quarterly by the
National Cancer Registrars Association
1330 Braddock Place, #520
Alexandria, VA 22314
(703) 299-6640
(703) 299-6620 FAX

Address any correspondence to:
National Cancer Registrars Association
1330 Braddock Place, #520
Alexandria, VA 22314

Address all editorial correspondence to:
Nadine Walker, MS, ODS-C
Email: JRMEditor@ncra-usa.org
Telephone: (703) 299-6640 ext. 327

Letters to the Editor
Letters to the Editor must be signed and include address and telephone number; none will be published anonymously. Letters subject to editing.

Editorial Policy
Opinions expressed in articles appearing in *Journal of Registry Management* are those of the authors and not necessarily those of the National Cancer Registrars Association, or the editor of *Journal of Registry Management*.

Accessing the *JRM*
Access to the *JRM* is available via NCRA's website at the *JRM* Library (ncra-usa.org). Full-text articles are also available on the National Library of Medicine's PubMed Central website.

Advertising Sales
Advertising is accepted with the understanding that products and services are professionally related and meet the ethical standards of public health practice. Acceptance by the *Journal* does not indicate or imply endorsement by the *Journal* or NCRA. Advertising sales should be directed to Marketing Coordinator, NCRA, 1330 Braddock Place, #520, Alexandria, VA 22314, (703) 299-6640, (703) 299-6620 FAX.

National Cancer Registrars Association

Journal of Registry Management

Winter 2025 • Volume 52 • Number 4

Contents, cont.

Letter from the Guest Editor

This is our sixth, annual NAACCR special edition of the *Journal of Registry Management.* The NAACCR special edition is an opportunity for the NAACCR community and those working with cancer registry data to publish cancer surveillance research, experience, or ideas. This special edition intends to showcase original research articles as well as short reports, editorials, and registry-specific experiences. The submission deadline for next year is Tuesday, September 29, 2026.

This Winter 2025 issue contains four original research articles, one short report, five "How I Did It" articles, and the two winning posters from the NAACCR 2025 Annual Conference, held in Hartford in June. Published articles undergo a peer review process with reviewers selected from NAACCR members active in the work of the Research and Data Use Steering Committee. I am grateful to our reviewers who met short turnaround times in the middle of Call for Data activities and the December holidays! The posters published went through a separate peer review process to be accepted at the NAACCR annual conference, met the criteria to be judged as part of our annual conference proceedings, and won their respective categories.

This volume contains articles on a variety of population-based cancer registry topics, including the impact of persistent poverty on cancer outcomes (Maguire, et al), the estimation of the cancer burden of hereditary-syndrome-related cancers (Morales, et al), trends in colorectal cancer incidence by age and subtype (Chung, et al), and an evaluation of the performance of different language models to extract treatment data from clinical notes (Shivanna, et al). The hereditary-syndrome-related cancers study is the subject of the CE quiz for this issue.

The "How I Did It" category topics include a resource from our colleagues in Australia for developing online statistics tools (Retell, et al), a status report on the development of a cancer data platform to support geospatial analysis (Rose, et al), a report on how to use the NCI-supported File*Pro to automate processing out of state cases (Jain, et al), an approach to assessing the quality of geocoded cancer registry data (Lefante, et al), and an early assessment of the linkage of Medicaid data with a non-SEER registry to assess cancer outcomes (Koroukian, et al). The short report discusses the complications of central sequence numbers that approach 60 for malignant tumors (Moore, et al).

The winning posters were presented and judged at our annual conference in Hartford, Connecticut. The Research and Data Use winning poster, from the Iowa Cancer Registry, presents a geospatial analysis of stage at diagnosis for HPV-related cancers (Janio, et al). The Standards and Registry Operations winning poster, from the New Jersey Stat Cancer Registry, covers case ascertainment of melanoma to support research (Flores, et al).

Please note that the opinions, findings, and conclusions in this report are those of the authors and do not necessarily represent the views of the NAACCR, NCRA, or the JRM.

I am grateful to continue our collaboration with NCRA and JRM on this special publication of NAACCR-focused articles. And, again, I would like to express my appreciation for both the article and poster reviewers.

Respectfully,
Recinda Sherman, PhD, MPH, ODS-C
Guest Editor, *JRM*

Original Article

The Effect of Persistent Poverty on Cancer Outcomes in California

Frances B. Maguire[a]; Ani S. Movsisyan Vernon[a]; Brenda M. Hofer[a]; Arti Parikh-Patel[a]; Theodore Wun[a,b]; Shehnaz K. Hussain[c]; Theresa H. M. Keegan[a,b]

Abstract: Background: Patients diagnosed with cancer in persistent poverty areas (PPAs) are more likely to present with advanced-stage disease and experience decreased survival compared to those in more affluent regions nationally. However, prior studies have not considered the independent effects of PPAs on stage at diagnosis and survival in California, which could inform efforts to improve cancer outcomes. Methods: We identified patients (≥15 years) newly diagnosed with nine cancers (female breast, prostate, lung, colorectal, cutaneous melanoma, oropharyngeal, liver, stomach, and cervical) from 2006 to 2020 in the California Cancer Registry. We used multivariable logistic regression to examine the association between living in PPAs (defined as having a poverty rate ≥20% for approximately 30 continuous years) and late-stage diagnosis (regional/distant versus in situ/localized). We used multivariable Cox proportional hazards regression to evaluate the effect of PPA on overall survival. Results: Among the 1,706,573 patients in the cohort, 6.0% (101,820) resided in PPAs. More patients in PPAs were younger, non-Hispanic Black or Hispanic, had public or no insurance, and lived in rural areas. Those living in PPAs (versus non-PPAs) had greater odds of late-stage diagnoses for all cancer types (adjusted odds ratios ranged from 1.09 to 1.46, $p<0.02$), except stomach. Across all cancer types, patients in PPAs had worse survival, even after considering stage at diagnosis and other factors (adjusted hazard ratios ranged from 1.05 to 1.35, $p>0.01$). Conclusion: Our findings of worse outcomes for patients living in PPAs highlight the need for more targeted interventions, such as education about cancer screening and funding to support improved access to healthcare in these areas.

Key words: *California, cancer, poverty*

Introduction

Although overall cancer incidence and mortality have declined in California since 1981, progress has not been equal for all populations. Disparities in cancer burden and outcomes by socioeconomic status, race/ethnicity, and geographic area persist, with worse cancer outcomes in minority, rural, and low income socioeconomic groups.[2–5] These disparities are pronounced in persistent poverty areas (PPAs), defined as areas where at least 20% of the population has lived in poverty for four consecutive time periods, 10 years apart, and persisting for approximately 30 years.[6,7] Based on persistent poverty classification at the census tract level, 60% of counties in California have PPAs ranging from 1% to 42% of census tracts.[8]

Stage at diagnosis is one of the strongest predictors of cancer survival.[9] Our prior work showed that fewer patients living in PPAs were diagnosed at localized stages[10] than patients in non-PPAs, suggesting barriers to accessing health care and cancer screenings. However, our prior work did not evaluate the independent effect of living in PPAs on stage at diagnosis and associations with survival.[10] US studies have found later stage at diagnosis and higher cancer mortality in patients living in PPAs, with cancer deaths in PPAs approximately 7% higher than in non-PPAs.[11,12] Understanding and addressing inequities faced by residents in PPAs is a National Cancer Institute (NCI) priority through its Persistent Poverty Initiative (PPI).[13] The NCI PPI has developed a framework that illustrates multilevel factors associated with PPAs that may contribute to developing cancer with associated potential interventions.[13] Quantifying the effect of living in PPAs on stage at cancer diagnosis and associations with survival in a large state with substantial geographic variation in poverty is a critical first step that can guide cancer control efforts and interventions, including addressing risk factors identified in the PPI, such as smoking and poor access to screening,[13] and developing tailored cancer education.[13] We therefore sought to examine the independent effect of persistent

[a]California Cancer Reporting and Epidemiologic Surveillance Program, University of California Davis Comprehensive Cancer Center [b]Center for Oncology Hematology Outcomes Research and Training (COHORT) and Division of Hematology and Oncology, University of California Davis School of Medicine [c]Department of Public Health Sciences, University of California Davis Health

Shehnaz K. Hussain and Theresa H. M. Keegan are co-senior authors.

Address correspondence to Frances B. Maguire, 1631 Alhambra Blvd., Suite 200, Sacramento, CA 95816, fbmaguire@ucdavis.edu

This work was supported by the UC Davis Comprehensive Cancer Center (P30CA093373), UC Davis Health System, under Orcid ID 0000-0001-8490-6229. The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; the Centers for Disease Control and Prevention's (CDC's) National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology, and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, the California Department of Public Health, the National Cancer Institute, or the Centers for Disease Control and Prevention or their Contractors and Subcontractors.

poverty on stage at diagnosis and survival for nine cancers that either have screening tests available or are associated with smoking.[14]

Methods

Study population

We identified patients in the California Cancer Registry (CCR) who were ≥15 years of age and diagnosed during the years 2006 to 2020 with in situ or invasive cancers of the breast (female), prostate, lung, colon and rectum (colorectal), skin (cutaneous melanoma), oropharynx, liver, stomach, or cervix, using the International Classification of Diseases for Oncology, 3rd edition site and histology codes (ICD-O-3), using the following Surveillance, Epidemiology, and End Results (SEER) site recodes:[15] 26000, 28010, 22030, 21041, 21042, 21043, 21044, 21045, 21046, 21047, 21048, 21049, 21051, 21052, 25010, 20010, 20020, 20030, 20040, 20050, 20060, 20070, 20080, 20090, 20100, 21071, 21072, 21020, and 27010. The CCR is a population-based cancer surveillance system that collects incidence reports on more than 190,000 cases of cancer diagnosed annually in California (99% of all diagnoses). It has collected data on tumor characteristics, treatment, and patient demographics since 1988. CCR's regional registries are supported by the NCI SEER program.

Sociodemographic and clinical characteristics

The CCR provided patient characteristics including the following demographic and clinical information: age at diagnosis, sex, health insurance type, race/ethnicity, rural/urban residence, comorbidity score, stage at diagnosis, marital status, and NCI designation of treating facility. We categorized patients into male or female sex and grouped health insurance into private (HMO, PPO, fee for service, Tricare, military, Veterans Affairs, Medicare with supplement), public/uninsured (Medicaid, Medicare without supplement, Medicare with Medicaid eligibility, Indian/Public Health Service, county funded, not insured), and unknown. We combined no insurance with public insurance because prior work has shown that patients with public insurance at diagnosis were uninsured or intermittently insured with public insurance before their diagnosis and that the small number of uninsured patients likely reflects retroactive Medicaid enrollment.[16] We grouped race/ethnicity into non-Hispanic (NH) White, NH Black, Hispanic, Asian/Pacific Islander, American Indian, and other/unknown, based on the North American Association of Central Cancer Registries' Hispanic and Asian/Pacific Islander identification algorithm.[17] Hispanic identification is not explicitly excluded from the Asian/Pacific Islander or American Indian categories because a logic is applied where these groups are categorized first from the race fields. Rural/urban residence was based on Medical Service Study Area (MSSA) designations developed by the Department of Health Care Access and Information and based on the U.S. census.[18] MSSAs were designed to identify medically underserved areas and are routinely used in California.[19] Patients with a frontier designation were combined into the rural category. Patients were classified into comorbidity categories of 0, 1, and 2 or more, using a Charlson comorbidity index based on 16 medical conditions, excluding cancer diagnoses, reported in the Department of Health Care Access and Information (HCAI) hospital discharge data linked to CCR data.[20] Stage at diagnosis was categorized as early (in situ, localized), late (regional, distant), and unknown using SEER summary stage. Treatment at NCI-designated cancer centers was determined by reviewing all abstracts from reporting facilities where patient treatment occurred and assigning NCI status if the cancer had been diagnosed when the facility had NCI designation.

PPA was defined by the U.S. Department of Agriculture's Economic Research Service and the National Cancer Institute as areas where ≥20% of residents were living in poverty over multiple decades (approximately 30 years) as determined by the 1990 and 2000 decennial censuses and the 2007–2011 and 2015–2019 American Community Survey 5-year estimates.[6] Data regarding the persistent poverty status for each census tract was accessed from the University of California San Francisco Health Atlas project.[21] Cancer patients were mapped to each census tract based on their address at the time of cancer diagnosis. We excluded patients with unknown addresses at the time of diagnosis (N=7,903, 0.5%).

Statistical Analysis

We used descriptive statistics to examine unadjusted associations between sociodemographic and clinical characteristics by PPA residence.

We used multivariable logistic regression models to evaluate the association between residing in a PPA and late stage of diagnosis versus early stage by cancer types. Models included the following sociodemographic and clinical factors that could impact access to care:[22] age, sex, health insurance type, race/ethnicity, rural/urban residence, and marital status. Results are presented as adjusted odds ratios (OR) and their associated 95% confidence intervals (CI). We used multivariable Cox proportional hazards regression to evaluate the effect of residing in a PPA on overall survival, adjusting for factors that can impact survival,[23] including age, sex, stage, comorbidity, health insurance type, race/ethnicity, rural/urban residence, marital status, and NCI-designation of treating facility. Survival time was calculated as days from the date of diagnosis to the date of death from any cause or the date of last follow-up, through 2022. We assessed proportional hazards assumptions by inspecting the survival curves [survival function versus survival time and log (-log) of the survival function versus the log of time]. Results are presented as adjusted hazard ratios (HR) and their associated 95% CIs. For all models (OR, HR), the stage groupings for each cancer type were evaluated using Kaplan Meier survival curves and, based on the results, unknowns (2–13%) were combined with the stage group they most aligned with (late-stage for all cancer types except prostate and cutaneous melanoma). Analyses were conducted using SAS software version 9.4 (SAS Institute Inc., Cary, North Carolina). All analyses were overseen by the Institutional Review Board of the University of California, Davis.

Results

Of the 1,706,573 cancers of the breast, prostate, lung, colon and rectum, skin (cutaneous melanoma), oropharynx, liver, stomach, and cervix diagnosed from 2006–2020, 6.0% (101,820) were diagnosed in patients residing in PPAs. Individuals with cancer residing in PPAs were younger (49.4% versus 44.9% 15–64 years), NH Black (17% versus 5.7%), Hispanic (37.1% versus 15.2%), had public or no insurance (52.8% versus 31.1%), lived in rural areas (18.4% versus 13.9%), and were diagnosed with distant stage disease (21.4% versus 15.3%) (Table 1).

Individuals living in PPAs (versus non-PPAs) had greater odds of late-stage diagnosis for all cancer types except stomach, ranging from 9% greater odds for oropharyngeal (OR 1.09, 95% CI 1.01,1.17) and liver (OR 1.09, 95% CI 1.03,1.15) cancer patients to 46% greater odds for patients with cutaneous melanoma (OR 1.46, 95% CI 1.34,1.59) (Figure 1). Factors associated with late-stage diagnoses included male sex, public or no insurance (versus private), and being unmarried. Increasing age was associated with greater odds of late-stage diagnosis for prostate and cervical cancers. The odds of late-stage diagnoses were greater for most racial/ethnic groups (relative to NH Whites) across multiple cancer types, with a few notable exceptions. Hispanic and Asian/Pacific Islander patients had lower odds of late-stage diagnosis for liver cancer and Asian/Pacific Islander patients had lower odds of late-stage diagnosis of breast cancers. Rural residents had greater odds of late-stage diagnosis for most cancer types (breast, prostate, lung, cutaneous melanoma, liver) (Supplemental Table 1).

Individuals living in PPAs had poorer survival than those living in non-PPAs for all cancer types (Figure 2, Supplemental Table 2). Patients residing in PPAs had increased risks of death relative to those living in non-PPAs, ranging from 5% for those with stomach (HR 1.05, 95% CI 1.01,1.09) cancer to 34% for those with cutaneous melanoma (HR 1.34, 95% CI 1.28,1.41). Five-year overall survival was 4–13% lower in PPAs compared to non-PPAs (Supplemental Table 3). Factors associated with poorer survival included male sex, increasing age, late-stage diagnosis, higher comorbidity, public or no insurance, rural location, being unmarried, and treatment at non-NCI-designated cancer centers. Hispanic and Asian/Pacific Islander patients (versus NH White) had better survival for most cancers while NH Black patients had poorer survival for most cancers (Supplemental Table 2).

Table 1. Characteristics of patients diagnosed with 9 cancers according to residence in a persistent poverty area (PPA), California, 2006–2020

Variables	Total n=1 706 573	Not residing in a PPA n=1 604 753 (94.0%)	Residing in a PPA n=101 820 (6.0%)
Age (years)	n (%)	n (%)	n (%)
15–39	56 741 (3.3)	52 489 (3.3)	4252 (4.2)
40–64	714 058 (41.8)	668 057 (41.6)	46 001 (45.2)
65–79	669 926 (39.3)	632 213 (39.4)	37 713 (37.0)
80 plus	265 848 (15.6)	251 994 (15.7)	13 854 (13.6)
Sex			
Male	825 341 (48.4)	774 516 (48.3)	50 825 (49.9)
Female	881 232 (51.6)	830 237 (51.7)	50 995 (50.1)
Race/ethnicity			
NH White	1 070 195 (62.7)	1 036 821 (64.6)	33 374 (32.8)
NH Black	108 339 (6.3)	91 048 (5.7)	17 291 (17.0)
Hispanic	281 295 (16.5)	243 497 (15.2)	37 798 (37.1)
Asian/PI	190 096 (11.1)	179 291 (11.2)	10 805 (10.6)
American Indian	9373 (0.5)	8436 (0.5)	937 (0.9)
Other/unknown	47 275 (2.8)	45 660 (2.8)	1615 (1.6)
Health insurance			
Private/military	1 011 789 (59.3)	971 596 (60.5)	40 193 (39.5)
Public/uninsured	553 468 (32.4)	499 751 (31.1)	53 717 (52.8)
Unknown	141 316 (8.3)	133 406 (8.3)	7910 (7.8)
Rural/urban residence			
Rural	241 403 (14.1)	222 694 (13.9)	18 709 (18.4)
Urban	1 465 168 (85.9)	1 382 058 (86.1)	83 110 (81.6)
Unknown	2 (0.0)	1 (0.0)	1 (0.0)

Table 1, *cont*. Characteristics of patients diagnosed with 9 cancers according to residence in a persistent poverty area (PPA), California, 2006–2020

Variables	Total n=1 706 573	Not residing in a PPA n=1 604 753 (94.0%)	Residing in a PPA n=101 820 (6.0%)
Stage at diagnosis			
In situ/localized	980 711 (57.5)	933 285 (58.2)	47 426 (46.6)
Regional	358 004 (21.0)	333 792 (20.8)	24 212 (23.8)
Distant	267 312 (15.7)	245 531 (15.3)	21 781 (21.4)
Unknown	100 546 (5.9)	92 145 (5.7)	8401 (8.3)
NCI Cancer Center			
No	1 475 975 (86.5)	1 388 226 (86.5)	87 749 (86.2)
Yes	230 598 (13.5)	216 527 (13.5)	14 071 (13.8)
Marital status			
Not married	619 056 (36.3)	567 867 (35.4)	51 189 (50.3)
Married	903 097 (52.9)	861 717 (53.7)	41 380 (40.6)
Unknown	184 420 (10.8)	175 169 (10.9)	9251 (9.1)
Cancer type			
Breast (female)	475 751 (27.9)	450 136 (28.1)	25 615 (25.2)
Prostate	316 139 (18.5)	298 037 (18.6)	18 102 (17.8)
Lung	256 908 (15.1)	238 609 (14.9)	18 299 (18.0)
Colorectal	229 053 (13.4)	212 375 (13.2)	16 678 (16.4)
Cutaneous melanoma	238 417 (14.0)	232 925 (14.5)	5492 (5.4)
Oral	64 300 (3.8)	60 173 (3.7)	4127 (4.1)
Liver	59 099 (3.5)	52 811 (3.3)	6288 (6.2)
Stomach	44 799 (2.6)	40 347 (2.5)	4452 (4.4)
Cervical	22 107 (1.3)	19 340 (1.2)	2767 (2.7)
Vital status			
Dead	661 301 (38.8)	610 806 (38.1)	50 495 (49.6)
Alive	1 045 272 (61.2)	993 947 (61.9)	51 325 (50.4)

Excludes 7903 patients missing PPA residence
Excludes 1039 patients <15 years of age
Abbreviations: NH, non-Hispanic; PI, Pacific Islander; NCI, National Cancer Institute

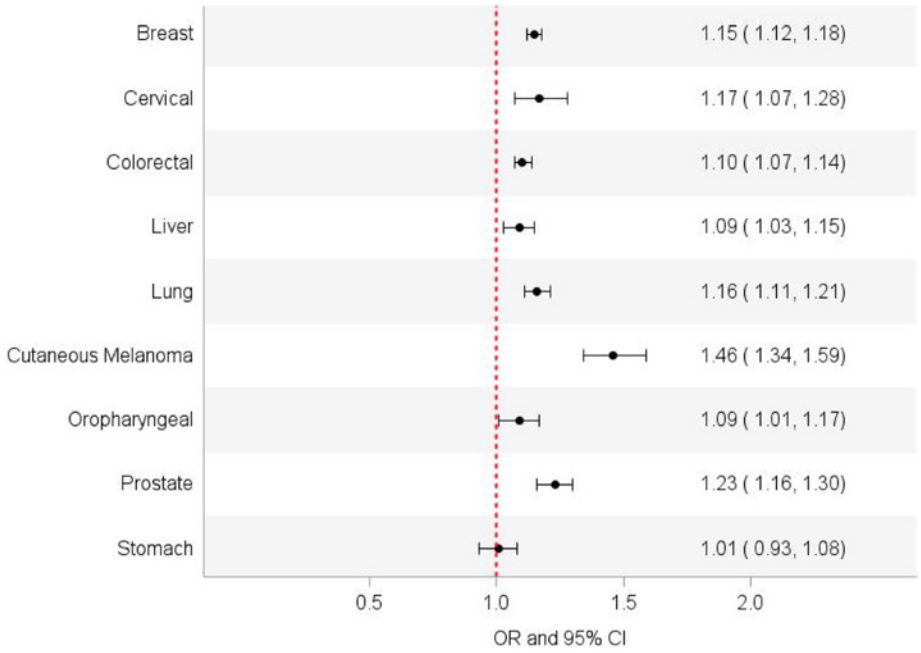
Discussion

In our population-based study of over 1,700,000 individuals in California, those living in PPAs were more likely to have worse survival across all 9 types of cancer and be diagnosed with late-stage breast, prostate, lung, colorectal, cutaneous melanoma, oropharyngeal, liver, and cervical cancers. Seven of the common cancers considered in this study have screening tests (breast, cervical, colorectal, lung, cutaneous melanoma, oral, prostate) that could potentially prevent the cancer (cervical, colorectal) or detect it at an early stage (breast, lung, cutaneous melanoma, oral, prostate), suggesting barriers to cancer screening in PPAs. Lower adherence to cancer screening[24–29] has been associated with later stage diagnosis and is identified in the NCI PPI conceptual framework as an important factor impacting cancer burden in PPAs.[13] In addition, even after adjustment for stage at diagnosis and other sociodemographic clinical factors, those residing in PPAs experienced worse survival, which may relate to barriers to access to high quality cancer care and cancer screening in these groups.

Our findings are consistent with prior work showing higher mortality rates and more late-stage diagnoses of cancer in PPAs compared to non-PPAs in the US.[7,12] Similar to these studies, we found increased odds of presenting with advanced stage disease and worse survival for lung, breast, and colorectal cancers, but also for prostate, cutaneous melanoma, oropharyngeal, and cervical cancers. Unlike previous studies, our work quantified the independent effect of PPAs on both stage at diagnosis and survival across 9 common cancers, data which can inform multilevel interventions and sustainable policies to reduce the cancer burden in these areas in California.[13]

The association of poverty with worse cancer outcomes and the contributions that individual factors play in this

Figure 1. Associations between residing in a persistent poverty area (yes vs. no) and late-stage (vs. early-stage) diagnosis by cancer type, California, 2006–2020



Abbreviations: OR, odds ratio; CI, confidence interval

Models adjusted for age, sex, health insurance type, race/ethnicity, rural/urban residence, and marital status. Stage at diagnosis defined as regional/distant (late-stage) and in situ/ localized (early-stage) from Surveillance, Epidemiology, and End Results (SEER) summary stage.

Supplemental Table 1. Characteristics associated with late-stage (vs. early-stage) diagnoses by cancer type

	Breast		Prostate		Lung		Colorectal	
Characteristics	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value
Persistent Poverty								
No (reference)								
Yes	1.15 (1.12, 1.18)	< 0.001	1.23 (1.16, 1.30)	< 0.001	1.16 (1.11, 1.21)	< 0.001	1.10 (1.07, 1.14)	< 0.001
Unknown	1.96 (1.76, 2.18)	< 0.001	1.06 (0.82, 1.36)	0.677	0.95 (0.76, 1.20)	0.680	1.41 (1.20, 1.64)	< 0.001
Age								
[15-39] (reference)								
[40-64]	0.50 (0.48, 0.51)	< 0.001	0.44 (0.26, 0.74)	0.002	1.59 (1.42, 1.79)	< 0.001	0.82 (0.78, 0.86)	< 0.001
[65-79]	0.34 (0.33, 0.35)	< 0.001	0.55 (0.32, 0.94)	0.027	1.11 (0.99, 1.25)	0.064	0.70 (0.66, 0.73)	< 0.001
[80 plus]	0.48 (0.46, 0.49)	< 0.001	2.47 (1.45, 4.21)	0.001	1.17 (1.04, 1.31)	0.008	0.86 (0.81, 0.90)	< 0.001
Sex								
Female (reference)								
Male	NA	NA	NA	NA	1.35 (1.32, 1.37)	< 0.001	1.01 (1.00, 1.03)	0.158
Health Insurance Status								
Private (reference)								
Public/uninsured	1.44 (1.42, 1.46)	< 0.001	1.64 (1.59, 1.69)	< 0.001	1.17 (1.14, 1.19)	< 0.001	1.28 (1.25, 1.30)	< 0.001
Unknown	2.51 (2.43, 2.60)	< 0.001	1.04 (0.96, 1.13)	0.339	4.89 (4.57, 5.24)	< 0.001	2.95 (2.81, 3.11)	< 0.001
Race/Ethnicity								
NH White (reference)								
NH Black	1.26 (1.22, 1.29)	< 0.001	1.32 (1.26, 1.39)	< 0.001	1.22 (1.17, 1.27)	< 0.001	0.98 (0.95, 1.02)	0.328
Hispanic	1.21 (1.19, 1.23)	< 0.001	1.24 (1.19, 1.29)	< 0.001	1.23 (1.19, 1.27)	< 0.001	1.06 (1.03, 1.08)	< 0.001
Asian/PI	0.94 (0.93, 0.96)	< 0.001	1.17 (1.12, 1.24)	< 0.001	1.22 (1.18, 1.26)	< 0.001	1.04 (1.02, 1.07)	0.001
American Indian	1.15 (1.06, 1.25)	0.001	1.24 (1.02, 1.51)	0.033	1.07 (0.94, 1.22)	0.307	1.05 (0.94, 1.17)	0.352
Other	1.01 (0.95, 1.08)	0.735	0.32 (0.27, 0.38)	< 0.001	0.6 (0.51, 0.72)	< 0.001	0.46 (0.43, 0.50)	< 0.001
Residence								
Urban (reference)								
Rural	1.07 (1.05, 1.09)	< 0.001	1.04 (1.00, 1.08)	0.044	1.12 (1.09, 1.15)	< 0.001	1.02 (0.99, 1.04)	0.203
Marital Status								
Married (reference)								
Not married	1.16 (1.15, 1.18)	< 0.001	1.78 (1.73, 1.84)	< 0.001	1.12 (1.09, 1.14)	< 0.001	1.17 (1.15, 1.19)	< 0.001
Unknown	1.26 (1.22, 1.29)	< 0.001	0.60 (0.56, 0.64)	< 0.001	1.12 (1.06, 1.18)	< 0.001	0.66 (0.64, 0.69)	< 0.001

Abbreviations: OR, odds ratio; CI, confidence interval; NH, non-Hispanic

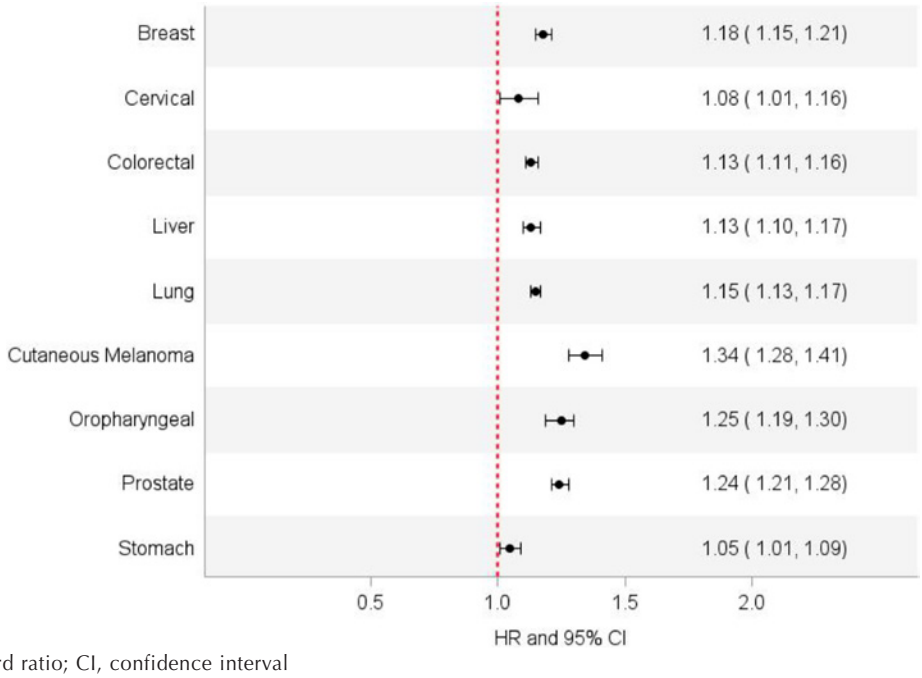
Models adjusted for all variables in the table. Stage at diagnosis defined as regional/distant (late-stage) and in situ/localized (early-stage) from Surveillance, Epidemiology, and End Results (SEER) summary stage.

Supplemental Table 1. Characteristics associated with late stage (vs. early stage) diagnoses by cancer type, continued

	Cutaneous Melanoma		Oral		Liver		Stomach		Cervix	
Characteristics	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value
Persistent Poverty										
No (reference)										
Yes	1.46 (1.34, 1.59)	< 0.001	1.09 (1.01, 1.17)	0.023	1.09 (1.03, 1.15)	0.002	1.01 (0.93, 1.08)	0.886	1.17 (1.07, 1.28)	< 0.001
Unknown	0.96 (0.68, 1.37)	0.830	1.58 (1.22, 2.05)	< 0.001	1.57 (1.14, 2.16)	0.005	1.54 (1.02, 2.32)	0.040	3.22 (1.74, 5.95)	< 0.001
Age										
[15-39] (reference)										
[40-64]	0.97 (0.91, 1.04)	0.423	2.17 (2.00, 2.35)	< 0.001	0.61 (0.52, 0.70)	< 0.001	0.71 (0.63, 0.81)	< 0.001	1.98 (1.85, 2.11)	< 0.001
[65-79]	0.81 (0.76, 0.87)	< 0.001	1.66 (1.52, 1.80)	< 0.001	0.61 (0.53, 0.71)	< 0.001	0.54 (0.48, 0.62)	< 0.001	3.74 (3.40, 4.12)	< 0.001
[80 plus]	0.95 (0.89, 1.03)	0.200	1.34 (1.22, 1.46)	< 0.001	0.73 (0.63, 0.86)	< 0.001	0.52 (0.45, 0.59)	< 0.001	5.03 (4.30, 5.89)	< 0.001
Sex										
Female (reference)										
Male	1.53 (1.47, 1.58)	< 0.001	1.99 (1.91, 2.06)	< 0.001	1.08 (1.04, 1.12)	< 0.001	1.38 (1.32, 1.44)	< 0.001	NA	NA
Health Insurance Status										
Private (reference)										
Public/uninsured	1.40 (1.34, 1.45)	< 0.001	1.25 (1.21, 1.30)	< 0.001	1.10 (1.06, 1.14)	< 0.001	1.10 (1.05, 1.15)	< 0.001	1.78 (1.68, 1.89)	< 0.001
Unknown	0.37 (0.34, 0.41)	< 0.001	2.06 (1.90, 2.25)	< 0.001	8.06 (7.30, 8.90)	< 0.001	2.87 (2.52, 3.26)	< 0.001	3.30 (2.81, 3.88)	< 0.001
Race/Ethnicity										
NH White (reference)										
NH Black	3.46 (2.77, 4.31)	< 0.001	1.49 (1.36, 1.62)	< 0.001	1.02 (0.96, 1.10)	0.503	0.93 (0.86, 1.01)	0.099	1.24 (1.10, 1.41)	0.001
Hispanic	2.42 (2.28, 2.57)	< 0.001	1.14 (1.08, 1.20)	< 0.001	0.94 (0.90, 0.98)	0.003	1.24 (1.18, 1.31)	< 0.001	0.97 (0.90, 1.03)	0.309
Asian/PI	3.55 (3.13, 4.02)	< 0.001	1.3 (1.23, 1.38)	< 0.001	0.86 (0.82, 0.90)	< 0.001	0.98 (0.93, 1.04)	0.490	1.10 (1.01, 1.20)	0.033
American Indian	1.51 (1.17, 1.94)	0.001	1.27 (1.02, 1.59)	0.035	1.00 (0.86, 1.17)	0.973	0.86 (0.65, 1.13)	0.281	1.11 (0.82, 1.49)	0.500
Other	0.19 (0.16, 0.23)	< 0.001	0.69 (0.60, 0.79)	< 0.001	0.85 (0.63, 1.16)	0.318	0.59 (0.45, 0.76)	< 0.001	0.65 (0.49, 0.86)	0.003
Residence										
Urban (reference)										
Rural	1.18 (1.13, 1.23)	< 0.001	1.01 (0.96, 1.06)	0.671	1.07 (1.02, 1.12)	0.01	0.89 (0.84, 0.96)	0.001	1.07 (0.98, 1.16)	0.131
Marital Status										
Married (reference)										
Not married	1.52 (1.47, 1.58)	< 0.001	1.28 (1.23, 1.33)	< 0.001	1.05 (1.01, 1.09)	0.008	1.11 (1.06, 1.16)	< 0.001	1.33 (1.25, 1.41)	< 0.001
Unknown	0.25 (0.23, 0.27)	< 0.001	0.74 (0.69, 0.80)	< 0.001	1.04 (0.94, 1.14)	0.440	0.79 (0.72, 0.88)	< 0.001	0.98 (0.86, 1.12)	0.764

Abbreviations: OR, odds ratio; CI, confidence interval; NH, non-Hispanic
Models adjusted for all variables in the table. Stage at diagnosis defined as regional/distant (late-stage) and in situ/localized (early-stage) from Surveillance, Epidemiology, and End Results (SEER) summary stage.

Figure 2. Associations between residing in a persistent poverty area (yes vs. no) and overall survival by cancer type, California 2006–2020



Abbreviations: HR, hazard ratio; CI, confidence interval
Models adjusted for age, sex, stage, comorbidity, treatment at NCI-designated cancer center, health insurance type, race/ethnicity, rural/urban residence, and marital status.

Supplemental Table 2. Characteristics associated with overall survival by cancer type

	Breast		Prostate		Lung		Colorectal	
Characteristics	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value
Persistent Poverty								
No (reference)								
Yes	1.18 (1.15, 1.21)	< 0.001	1.24 (1.21, 1.28)	< 0.001	1.15 (1.13, 1.17)	< 0.001	1.13 (1.11, 1.16)	< 0.001
Unknown	1.38 (1.22, 1.54)	< 0.001	1.34 (1.21, 1.48)	< 0.001	0.49 (0.45, 0.54)	< 0.001	1.22 (1.08, 1.38)	0.001
Age								
[15-39] (reference)								
[40-64]	0.84 (0.81, 0.87)	< 0.001			1.63 (1.53, 1.75)	< 0.001	1.15 (1.10, 1.20)	< 0.001
[65-79]	1.52 (1.47, 1.58)	< 0.001	2.12 (2.08, 2.16)	< 0.001	1.87 (1.75, 2.01)	< 0.001	1.72 (1.64, 1.79)	< 0.001
[80 plus]	4.62 (4.45, 4.80)	< 0.001	6.34 (6.20, 6.49)	< 0.001	2.84 (2.66, 3.04)	< 0.001	3.47 (3.32, 3.62)	< 0.001
Sex								
Female (reference)								
Male	NA	NA	NA	NA	1.27 (1.26, 1.29)	< 0.001	1.19 (1.18, 1.21)	< 0.001
Stage								
Early (reference)								
Late	3.05 (3.02, 3.09)	< 0.001	5.95 (5.83, 6.07)	< 0.001	3.40 (3.35, 3.44)	< 0.001	2.74 (2.70, 2.78)	< 0.001
Unknown			2.82 (2.75, 2.89)	< 0.001				
Charlson Index								
0 (reference)								
1	1.45 (1.42, 1.47)	< 0.001	1.48 (1.44, 1.51)	< 0.001	1.17 (1.15, 1.18)	< 0.001	1.24 (1.22, 1.26)	< 0.001
2 or more	2.36 (2.32, 2.40)	< 0.001	2.50 (2.45, 2.55)	< 0.001	1.50 (1.48, 1.52)	< 0.001	1.83 (1.80, 1.86)	< 0.001
Unknown	1.38 (1.35, 1.41)	< 0.001	1.05 (1.04, 1.07)	< 0.001			1.27 (1.24, 1.29)	< 0.001
Health Insurance Status								
Private (reference)								
Public/uninsured	1.40 (1.39, 1.42)	< 0.001	1.26 (1.24, 1.28)	< 0.001	1.11 (1.10, 1.12)	< 0.001	1.30 (1.29, 1.32)	< 0.001
Unknown					2.16 (2.12, 2.20)	< 0.001		
Race/Ethnicity								
NH White (reference)								
Black	1.19 (1.16, 1.22)	< 0.001	1.09 (1.07, 1.12)	< 0.001	0.97 (0.96, 0.99)	0.0012	1.12 (1.10, 1.15)	< 0.001
Hispanic	0.90 (0.89, 0.92)	< 0.001	0.87 (0.85, 0.88)	< 0.001	0.97 (0.95, 0.98)	< 0.001	0.92 (0.91, 0.94)	< 0.001
Asian/PI	0.77 (0.75, 0.79)	< 0.001	0.81 (0.79, 0.84)	< 0.001	0.83 (0.82, 0.85)	< 0.001	0.85 (0.84, 0.87)	< 0.001
American Indian	1.03 (0.95, 1.12)	0.477	1.16 (1.05, 1.28)	0.0027	1.01 (0.96, 1.07)	0.579	0.97 (0.90, 1.05)	0.484
Other	0.27 (0.24, 0.31)	< 0.001	0.14 (0.13, 0.15)	< 0.001	0.53 (0.48, 0.60)	< 0.001	0.23 (0.21, 0.27)	< 0.001
Residence								
Urban (reference)								
Rural	1.07 (1.05, 1.09)	< 0.001	1.08 (1.05, 1.10)	< 0.001	1.07 (1.05, 1.08)	< 0.001	1.05 (1.03, 1.06)	< 0.001
Marital Status								
Married (reference)								
Not married	1.39 (1.37, 1.41)	< 0.001	1.47 (1.44, 1.49)	< 0.001	1.23 (1.22, 1.24)	< 0.001	1.31 (1.29, 1.32)	< 0.001
Unknown	1.37 (1.34, 1.42)	< 0.001	0.98 (0.96, 1.00)	0.101	0.94 (0.92, 0.97)	< 0.001	1.02 (0.99, 1.05)	0.144
Treatment at NCI-designated center								
Yes (reference)								
No	1.28 (1.25, 1.31)	< 0.001	1.42 (1.39, 1.46)	< 0.001	1.42 (1.40, 1.44)	< 0.001	1.17 (1.15, 1.20)	< 0.001

Abbreviations: HR, hazard ratio; CI, confidence interval; NH, non-Hispanic; NCI, National Cancer Institute

Models are adjusted for all variables in the table.

Some categories were combined for model fit: age: 15–39 and 40–64 combined for prostate cancer and stomach cancer; comorbidity: unknown combined with 1 for lung cancer; health insurance: unknown combined with public for breast, prostate, colorectal, stomach, and cervix cancers.

Supplemental Table 2. Characteristics associated with overall survival by cancer type, continued

Characteristics	Cutaneous Melanoma		Oral		Liver		Stomach		Cervix	
	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value
Persistent Poverty										
No (reference)										
Yes	1.34 (1.28, 1.41)	< 0.001	1.25 (1.19, 1.30)	< 0.001	1.13 (1.10, 1.17)	< 0.001	1.05 (1.01, 1.09)	0.007	1.08 (1.01, 1.16)	< 0.001
Unknown	1.79 (1.62, 1.97)	< 0.001	1.09 (0.91, 1.29)	0.356	0.78 (0.66, 0.93)	0.004	1.13 (0.92, 1.38)	0.246	1.21 (0.79, 1.84)	0.376
Age										
[15-39] (reference)										
[40-64]	1.94 (1.80, 2.10)	< 0.001	1.90 (1.73, 2.08)	< 0.001	1.36 (1.25, 1.49)	< 0.001			1.37 (1.28, 1.47)	< 0.001
[65-79]	4.80 (4.46, 5.18)	< 0.001	3.01 (2.75, 3.30)	< 0.001	1.54 (1.41, 1.68)	< 0.001	1.11 (1.08, 1.14)	< 0.001	1.98 (1.83, 2.14)	< 0.001
[80 plus]	15.00 (13.91, 16.17)	< 0.001	6.40 (5.83, 7.03)	< 0.001	2.36 (2.16, 2.59)	< 0.001	1.91 (1.85, 1.96)	< 0.001	3.80 (3.46, 4.17)	< 0.001
Sex										
Female (reference)										
Male	1.47 (1.45, 1.50)	< 0.001	1.10 (1.07, 1.13)	< 0.001	1.09 (1.07, 1.11)	< 0.001	1.11 (1.08, 1.14)	< 0.001	NA	NA
Stage										
Early (reference)										
Late	4.00 (3.91, 4.10)	< 0.001	1.95 (1.90, 2.00)	< 0.001	2.46 (2.41, 2.51)	< 0.001	3.47 (3.37, 3.57)	< 0.001	4.91 (4.62, 5.22)	< 0.001
Unknown										
Charlson Index										
0 (reference)										
1	1.49 (1.45, 1.53)	< 0.001	1.51 (1.46, 1.56)	< 0.001	1.00 (0.97, 1.04)	0.958	1.08 (1.04, 1.11)	< 0.001	1.35 (1.26, 1.44)	< 0.001
2 or more	2.19 (2.14, 2.25)	< 0.001	1.99 (1.92, 2.05)	< 0.001	1.35 (1.31, 1.38)	< 0.001	1.42 (1.38, 1.46)	< 0.001	2.29 (2.15, 2.44)	< 0.001
Unknown	0.82 (0.80, 0.84)	< 0.001	1.04 (1.00, 1.08)	0.026	0.98 (0.95, 1.01)	0.253	1.13 (1.09, 1.18)	< 0.001	1.07 (1.00, 1.14)	0.041
Health Insurance Status										
Private (reference)										
Public/uninsured	1.16 (1.14, 1.18)	< 0.001	1.38 (1.35, 1.42)	< 0.001	1.13 (1.11, 1.15)	< 0.001	1.17 (1.15, 1.20)	< 0.001	1.32 (1.26, 1.38)	< 0.001
Unknown	1.24 (1.20, 1.28)	< 0.001	1.97 (1.87, 2.07)	< 0.001	2.67 (2.58, 2.76)	< 0.001				
Race/Ethnicity										
NH White (reference)										
Black	1.47 (1.27, 1.70)	< 0.001	1.26 (1.20, 1.33)	< 0.001	1.01 (0.98, 1.05)	0.526	0.95 (0.91, 0.99)	0.027	1.22 (1.12, 1.32)	< 0.001
Hispanic	1.19 (1.14, 1.24)	< 0.001	1.07 (1.03, 1.11)	< 0.001	0.97 (0.95, 0.99)	0.008	0.97 (0.95, 1.00)	0.044	0.83 (0.78, 0.87)	< 0.001
Asian/PI	1.32 (1.20, 1.45)	< 0.001	1.00 (0.96, 1.04)	0.891	0.88 (0.85, 0.90)	< 0.001	0.80 (0.77, 0.82)	< 0.001	0.78 (0.72, 0.83)	< 0.001
American Indian	1.12 (0.96, 1.30)	0.166	1.02 (0.89, 1.18)	0.759	0.97 (0.89, 1.05)	0.437	1.10 (0.94, 1.28)	0.240	0.84 (0.66, 1.07)	0.157
Other	0.11 (0.10, 0.12)	< 0.001	0.20 (0.16, 0.24)	< 0.001	0.48 (0.39, 0.59)	< 0.001	0.33 (0.25, 0.43)	< 0.001	0.17 (0.10, 0.30)	< 0.001
Residence										
Urban (reference)										
Rural	1.09 (1.06, 1.12)	< 0.001	1.03 (1.00, 1.07)	0.0423	1.08 (1.05, 1.11)	< 0.001	1.10 (1.06, 1.14)	< 0.001	1.00 (0.94, 1.07)	0.941
Marital Status										
Married (reference)										
Not married	1.42 (1.39, 1.45)	< 0.001	1.47 (1.43, 1.50)	< 0.001	1.17 (1.15, 1.19)	< 0.001	1.22 (1.19, 1.25)	< 0.001	1.16 (1.10, 1.22)	< 0.001
Unknown	1.07 (1.04, 1.10)	< 0.001	1.02 (0.97, 1.07)	0.580	0.96 (0.91, 1.01)	0.104	0.92 (0.87, 0.98)	0.007	1.02 (0.91, 1.15)	0.635
Treatment at NCI-designated center										
Yes (reference)										
No	1.09 (1.06, 1.12)	< 0.001	1.13 (1.10, 1.16)	< 0.001	1.60 (1.57, 1.64)	< 0.001	1.41 (1.37, 1.46)	< 0.001	1.37 (1.29, 1.44)	< 0.001

Abbreviations: HR, hazard ratio; CI, confidence interval; NH, non-Hispanic
Models are adjusted for all variables in the table.
Some categories were combined for model fit: age: 15–39 and 40–64 combined for prostate cancer and stomach cancer; comorbidity: unknown combined with 1 for lung cancer; health insurance: unknown combined with public for breast, prostate, colorectal, stomach and cervix cancers.

Supplemental Table 3. Stage, vital status, and survival by persistent poverty area designation and cancer site

Characteristic	Breast	Prostate	Lung	Colorectal	Cutaneous Melanoma	Oral	Liver	Stomach	Cervix
				PPA					
Stage at diagnosis (%, n)*									
Early (in situ/localized)	62.6% (16042)	66.2% (11987)	16.4% (3007)	36.5% (6081)	80.8% (4440)	27.5% (1136)	40.3% (2537)	26.1% (1162)	37.4% (1034)
Late (regional/distant)	34.1% (8746)	21.7% (3936)	74.9% (13711)	55.6% (9276)	12.2% (670)	65.2% (2690)	43.6% (2743)	60.8% (2705)	54.8% (1516)
Vital status (%, n)									
Dead	25.9% (6624)	33.8% (6122)	84.6% (15482)	53.1% (8848)	27.9% (1530)	55.4% (2286)	83.0% (5219)	73.5% (3270)	40.3% (1114)
Alive	74.1% (18991)	66.2% (11980)	15.4% (2817)	46.9% (7830)	72.1% (3962)	44.6% (1841)	17.0% (1069)	26.5% (1182)	59.7% (1653)
				Non-PPA					
Stage at diagnosis (%, n)**									
Early (in situ/localized)	70.4% (317006)	71.0% (211494)	20.9% (49938)	40.8% (86547)	88.3% (205602)	33.0% (19848)	43.0% (22709)	28.2% (11363)	44.9% (8678)
Late (regional/distant)	27.6% (124090)	20.8% (61948)	71.8% (171414)	53.1% (112767)	6.8% (15840)	61.0% (36577)	42.8% (22607)	60.3% (24349)	49.8% (9631)
Vital status (%, n)									
Dead	20.2% (91035)	24.9% (74209)	79.5% (189813)	48.6% (103188)	21.0% (48843)	44.4% (26711)	78.9% (41647)	71.2% (28710)	34.4% (6650)
Alive	79.8% (359101)	75.1% (223828)	20.5% (48796)	51.4% (109187)	79.0% (184082)	55.6% (33462)	21.1% (11164)	28.8% (11637)	65.6% (12690)
5-year overall survival (%)									
PPA	79.4	76.9	14.1	50.6	76.9	47.3	14.1	24.8	59.6
Non-PPA	85.8	84.2	21.2	57.3	84.2	60.5	19.6	28.6	66.5

*Excludes 8401 cases with unknown stage information
**Excludes 92,145 cases with unknown stage information

relationship are not well understood. Contributing factors include worse health behaviors (tobacco use, obesity, and physical inactivity)[30–32] and increased barriers to healthcare, such as lack of health insurance and shortages of health-care providers in PPAs.[33,34] We observed that the majority of patients in PPAs had public or no insurance coverage. Poor access to care results in insufficient cancer screening.[35,36] Finally, social determinants of health are widely recognized as playing a large role in worse cancer outcomes in PPAs and include economic and social policies, discrimination, violence, environmental pollutants, literacy, and political structures.[13,37]

In our study, the strongest associations with late-stage diagnoses and worse overall survival in PPAs were among patients with cutaneous melanoma. This could be attributable to a shortage of dermatologists and a lack of dermatologic education among primary care providers in PPAs compared to non-PPAs.[38,39] Individuals with prostate cancer living in PPAs also had strong associations with both late-stage diagnosis and worse survival. Prior work has shown lower prostate cancer screening rates and less definitive treatment among individuals diagnosed with prostate cancer in PPAs.[40,41] The lack of association between PPAs and late-stage diagnosis for patients with stomach cancer is not consistent with prior work and warrants additional follow up research.[42]

This study had limitations. The poverty measure we used includes census tracts with large off-campus college student populations, which are transient resident groups and could result in misclassification of poverty status.[43] An analysis evaluating the impact of excluding off-campus college students on poverty rates found that the inclusion of off-campus college students may have a strong impact on determination of poverty rates at finer levels of geography, such as the census tract level that was the geographical unit used to determine PPA status of participants in this study.[44] Another limitation was our inability to include smoking status in our models due to the high level of missing values for this variable during the time period of this study. We also lacked other data, such as cancer screening and health behaviors, that could impact stage at diagnosis and survival. Despite these limitations, this study used high-quality population-based data from a large ethnically diverse state to report the independent effect of PPAs on outcomes for 9 cancer types. Additionally, evaluation of PPAs as a measure of economic and social disadvantage provides a broader look compared to a point in time measure of the barriers beyond the individual that perpetuate disparities.[45]

In conclusion, our study found significant disparities in stage at diagnosis and overall survival for cancer patients residing in PPAs. Our findings highlight the need for more targeted interventions, such as education about cancer screening and funding to support improved access to healthcare in PPAs. In addition, our results indicate that more research is needed to better understand barriers to accessing cancer care among residents of PPAs. Understanding the cancer outcomes among those living in PPAs is essential to guide public health strategies and policy initiatives.

Declarations

Conflicts of interest/competing interest:

There are no conflicts of interest to report.

Availability of data and material

The data that support the findings of this study are available from the California Cancer Registry. Access is granted through an application process by the management or data custodians (**https://www.ccrcal.org/retrieve-data/**).

References

1. California Cancer Registry, California Department of Public Health. CAL*Explorer: An interactive website for California Cancer Registry (CCR) statistics. 2025. **https://explorer.ccrcal.org/** (accessed September 21, 2025).
2. Kim G, Qin J, Hall CB, In H. Association between socioeconomic and insurance status and delayed diagnosis of gastrointestinal cancers. *J Surg Res* 2022;279:170–86.
3. Zahnd WE, Murphy C, Knoll M, Benavidez GA, Day KR, Ranganathan R, Luke P, Zgodic A, Shi K, Merrell MA, Crouch EL, Brandt HM, Eberth JM. The intersection of rural residence and minority race/ethnicity in cancer disparities in the United States. *Int J Environ Res Public Health* 2021;18(4).
4. Zavala VA, Bracci PM, Carethers JM, Carvajal-Carmona L, Coggins NB, Cruz-Correa MR, Davis M, de Smith AJ, Dutil J, Figueiredo JC, Fox R, Graves KD, Gomez SL, Llera A, Neuhausen SL, Newman L, Nguyen T, Palmer JR, Palmer NR, Pérez-Stable EJ, Piawah S, Rodriquez EJ, Sanabria-Salas MC, Schmit SL, Serrano-Gomez SJ, Stern MC, Weitzel J, Yang JJ, Zabaleta J, Ziv E, Fejerman L. Cancer health disparities in racial/ethnic minorities in the United States. *Br J Cancer* 2021;124(2):315–32.
5. Hofer BM, Maguire FB, Morris CR, Movsisyan AS, Parikh-Patel A, Kizer KW. Rural-urban variations in cancer incidence, detection, and survival in California. 2019. UC Davis Comprehensive Cancer Center. **https://health.ucdavis.edu/cancer/population-health/cancer-registry-data.html** (accessed December 18, 2023).
6. United States Department of Agriculture. Poverty Area Measures. 2022. Economic Research Service. **https://www.ers.usda.gov/data-products/poverty-area-measures/** (accessed August 18, 2024).
7. Moss JL, Pinto CN, Srinivasan S, Cronin KA, Croyle RT. Persistent poverty and cancer mortality rates: an analysis of county-level poverty designations. *Cancer Epidemiol Biomarkers Prev* 2020;29(10):1949–54.
8. University of California San Francisco School of Medicine. *UCSF Health Atlas*. 2022. Dean's Office of Population Health and Health Equity. **http://www.healthatlas.ucsf.edu** (accessed December 18, 2023).
9. American Cancer Society. Cancer Facts & Figures 2025. 2025. **https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2025/2025-cancer-facts-and-figures-acs.pdf** (accessed September 21, 2025).
10. Movsisyan AS, Maguire FB, Ullah AT, Hofer BM, Parikh-Patel A, Keegan THM, Wun T. *The Burden of Cancer in Persistent Poverty Areas of California*. 2023. California Cancer Reporting and Epidemiologic Surveillance (CalCARES) Program, UC Davis Comprehensive Cancer Center. **https://health.ucdavis.edu/cancer/population-health/cancer-registry-data.html** (accessed December 18, 2023).
11. Moss JL, Pinto CN, Srinivasan S, Cronin KA, Croyle RT. Enduring cancer disparities by persistent poverty, rurality, and race: 1990–1992 to 2014–2018. *J Nat Cancer Inst* 2022;114(6):829–36.
12. Papageorge MV, Woods AP, de Geus SWL, Ng SC, McAneny D, Tseng JF, Kenzik KM, Sachs TE. The persistence of poverty and its impact on cancer diagnosis, treatment and survival. *Ann Surg* 2023;277(6):995–1001.
13. Hiatt RA, Kennedy AE, Adams AK, Basen-Engquist KM, Bondy ML, Carvajal-Carmona LG, Fouad M, Grusky DB, Matthews P, McNeill LH, Phillips E, Pinheiro LC, Pisu M, Rechis R, Rehkopf DH, Tamimi RM, Terry MB, Wetter DW, Srinivasan S. The Persistent Poverty Initiative's framework to address social and economic determinants of cancer. *Cancer Epidemiol Biomarkers Prev* 2025,34(8):1235–9.

14. U.S. Department of Health & Human Services. *The Health Consequences of Smoking: 50 years of Progress: a Report of the Surgeon General*. 2014. U.S. Department of Health and Human Services, CDC. **https://www.hhs.gov/surgeongeneral/reports-and-publications/tobacco/index.html** (accessed October 12, 2025).
15. National Cancer Institute, Surveillance, Epidemiology, and End Results Program. Site Recode ICD-0-3/WHO 2008 Definition. 2008. **https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html** (accessed October 14, 2025).
16. Keegan THM, Parsons HM, Chen Y, Maguire FB, Morris CR, Parikh-Patel A, Kizer KW, Wun T. Impact of health insurance on stage at cancer diagnosis among adolescents and young adults. *J Nat Cancer Inst* 2019;111(11):1152–60.
17. North American Association of Central Cancer Registries. NAACCR Hispanic and Asian/Pacific Islander Identification Algorithm (NHAPIIA). 2019. **https://www.naaccr.org/analysis-and-data-improvement-tools/#NHAPIIA** (accessed March 15, 2024).
18. Department of Health Care Access and Information. Medical Service Study Area (MSSA) - Census Detail. 2014. California Open Data Portal. **https://data.chhs.ca.gov/dataset/medical-service-study-areas** (accessed January 21 2024).
19. California State Geoportal. Medical Service Study Area. 2024. **https://gis.data.ca.gov/search?q=mssa** (accessed January 12, 2026).
20. Lichtensztajn DY, Giddings BM, Morris CR, Parikh-Patel A, Kizer KW. Comorbidity index in central cancer registries: the value of hospital discharge data. *Clin Epidemiol*. 2017;9:601–9.
21. University of California San Francisco Population Health and Health Equity. UCSF Health Atlas. 2022. **https://healthatlas.ucsf.edu/?v%5B0%5D=pct_eli_all&v%5B1%5D=pct_disabled&f%5B0%5D=national&g=county&zoomgeo=true&y=2020&c%5B0%5D=-96.155&c%5B1%5D=39.4493&z=4.13&s=** (accessed September 18, 2024).
22. Agency for Health Care Research and Quality. 2021 National Healthcare Quality and Disparities Report, Access to Healthcare and Disparities in Access. 2021. **https://www.ncbi.nlm.nih.gov/books/NBK578537/** (accessed October 14, 2024).
23. Afshar N, English DR, Milne RL. Factors explaining socio-economic inequalities in cancer survival: a systematic review. *Cancer Control* 2021;28:10732748211011956.
24. Talham CJ, Montiel Ishino FA, O'Brien KM, Sandler DP, Williams F. Breast cancer screening among Hispanic and non-Hispanic White women by birthplace in the Sister Study. *Cancer Med* 2022;11(8):1913–22.
25. Sosa E, D'Souza G, Akhtar A, Sur M, Love K, Duffels J, Raz DJ, Kim JY, Sun V, Erhunmwunsee L. Racial and socioeconomic disparities in lung cancer screening in the United States: A systematic review. *CA Cancer J Clin* 2021;71(4):299–314.
26. McLeod MR, Galoosian A, May FP. Racial and ethnic disparities in colorectal cancer screening and outcomes. *Hematol Oncol Clin North Am* 2022;36(3):415–28.
27. Fuzzell LN, Perkins RB, Christy SM, Lake PW, Vadaparampil ST. Cervical cancer screening in the United States: challenges and potential solutions for underscreened groups. *Prev Med* 2021;144:106400.
28. Gawron AJ, Staub J, Bielefeldt K. Impact of health insurance, poverty, and comorbidities on colorectal cancer screening: insights from the Medical Expenditure Panel Survey. *Dig Dis Sci* 2021;66(1):70–7.
29. Zhao G, Okoro CA, Li J, Town M. Health insurance status and clinical cancer screenings among U.S. Adults. *Am J Prev Med* 2018;54(1):e11–e9.
30. Moss JL, Pinto CN, Shen C. Prevalence of cancer risk behaviors by county-level persistent poverty. *Cancer Epidemiol* 2025;94:102735.
31. Drope J, Liber AC, Cahn Z, Stoklosa M, Kennedy R, Douglas CE, Henson R, Drope J. Who's still smoking? Disparities in adult cigarette smoking prevalence in the United States. *CA Cancer J Clin* 2018;68(2):106–15.
32. Bennett KJ, Probst JC, Pumkam C. Obesity among working age adults: the role of county-level persistent poverty in rural disparities. *Health Place* 2011;17(5):1174–81.
33. Lin Y, Monnette A, Shi L. Effects of Medicaid expansion on poverty disparities in health insurance coverage. *Int J Equity Health* 2021;20(1):171.
34. Streeter RA, Snyder JE, Kepley H, Stahl AL, Li T, Washko MM. The geographic alignment of primary care Health Professional Shortage Areas with markers for social determinants of health. *PLoS One* 2020;15(4):e0231443.
35. Continelli T, McGinnis S, Holmes T. The effect of local primary care physician supply on the utilization of preventive health services in the United States. *Health Place* 2010;16(5):942–51.
36. Belasco EJ, Gong G, Pence B, Wilkes E. The impact of rural health care accessibility on cancer-related behaviors and outcomes. *Appl Health Econ Health Policy* 2014;12(4):461–70.
37. Centers for Disease Control and Prevention. Social Determinants of Health. 2024. **https://www.cdc.gov/public-health-gateway/php/about/social-determinants-of-health.html** (accessed September 29, 2025).
38. Najmi M, Brown AE, Harrington SR, Farris D, Sepulveda S, Nelson KC. A systematic review and synthesis of qualitative and quantitative studies evaluating provider, patient, and health care system-related barriers to diagnostic skin cancer examinations. *Arch Dermatol Res* 2022;314(4):329–40.
39. Feng H, Berk-Krauss J, Feng PW, Stein JA. Comparison of dermatologist density between urban and rural counties in the United States. *JAMA Dermatol* 2018;154(11):1265–71.
40. Baldwin LM, Andrilla CH, Porter MP, Rosenblatt RA, Patel S, Doescher MP. Treatment of early-stage prostate cancer among rural and urban patients. *Cancer* 2013;119(16):3067–75.
41. Weiner AB, Matulewicz RS, Tosoian JJ, Feinglass JM, Schaeffer EM. The effect of socioeconomic status, race, and insurance type on newly diagnosed metastatic prostate cancer in the United States (2004–2013). *Urol Oncol* 2018;36(3):91.e1–e6.
42. Greenlee RT, Howe HL. County-level poverty and distant stage cancer in the United States. *Cancer Causes Control* 2009;20(6):989–1000.
43. Benson C, Bishaw A. *Examining the Effect of Off-Campus College Students on Poverty Rates*. United States Census Bureau. 2017. **https://www.census.gov/newsroom/blogs/random-samplings/2017/12/off-campus.html** (accessed September 30, 2025).
44. Tippett R. Impact of off-campus college students on local poverty rates. Carolina Demography, The University of North Carolina at Chapel Hill. 2015. **https://carolinademography.cpc.unc.edu/2015/03/09/impact-of-off-campus-college-students-on-local-poverty-rates/** (accessed September 30, 2025).
45. United States Department of Agriculture, Economic Research Service. *Rural Poverty and Well-Being*. 2025. **https://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being** (accessed October 24, 2025).

Original Article

The Burden of Cancers Associated with Hereditary Breast and Ovarian Cancer Syndrome, 2015–2019: A Population-based Cancer Registry Analysis

Caroline B. Morales, MPH[a,b]; Lisa E. Paddock, PhD[a,b]; Katelyn Bleeker, MPH, ODS-C[a]; Antoinette M. Stroup, PhD[a,b]

Abstract: Background: Hereditary breast and ovarian cancer syndrome (HBOCS) is an inherited condition that increases the risk of developing several types of cancer; however, the true burden of HBOCS-related cancers remains unclear at the population level. Methods: We used 2015–2019 breast and ovarian cancer data (N=55,437) from the New Jersey State Cancer Registry to estimate the age-adjusted incidence of HBOCS-related breast and ovarian cancers in New Jersey, overall and by county. Cancers meeting 5 HBOCS diagnostic characteristics selected from the 2021 National Comprehensive Cancer Network risk assessment guidelines were flagged as possibly being HBOCS-related and referred to as "possible-HBOCS (pHBOCS)." We evaluated the racial/ethnic and stage patterns of HBOCS cancers and compared them to all cancers to further characterize the HBOCS burden. Age-adjusted incidence rates (AAIR) per 100,000 population, rate ratios (RR), and 95% confidence intervals (CI) were computed using the Surveillance, Epidemiology, and End Results Program (SEER) *Stat Database. Results: We identified 12,679 pHBOCS cancers among NJ residents from 2015–2019. The female AAIR of pHBOCS breast and ovarian cancers was 50.6 per 100,000 population. The rates of pHBOCS varied by county, and pHBOCS county rates varied by race. Among Hispanic women, the incidence was significantly higher in Hunterdon County (AAIR 115, 95% CI:68.4–181.1) than statewide (AAIR 40.6, 95% CI:38.7–42.6). The pHBOCS incidence was significantly lower for Asian/Pacific Islander (RR=0.85, 95% CI:0.80–0.91), Hispanic (RR=0.74, 95% CI:0.70–0.79), and non-Hispanic Black women (RR=0.90 95% CI:0.85–0.95) than non-Hispanic White women. Compared to local-stage ovarian cancers, the incidence of distant-stage was significantly higher in the overall cancer population (RR=2.59, 95% CI:2.37–2.83) and the pHBOCS cancers (RR=3.22, 95% CI:2.93–3.55). The rate of regional-stage ovarian cancers was only significantly higher than local for the pHBOCS group (RR=1.21, 95% CI:1.08–1.36). Conclusions: This study can be used as a starting point for building a framework that better estimates the population burden of HBOCS cancers. Quantifying the HBOCS impact improves our understanding of the true burden. Due to data availability limitations, true estimates could not be calculated. Policies supporting population-level, genetic/family history data collection are needed to accurately quantify HBOCS incidence.

Key words: *breast cancer, familial and hereditary cancers, HBOC, ovarian cancer*

Introduction

Breast and ovarian cancers are the second and sixth leading causes of cancer-related deaths among US women, respectively,[1] and nearly 530 men were projected to die from breast cancer in 2024.[2] Up to 10% of breast cancers and 15% of ovarian cancers have been attributed to inherited mutations,[3,4] suggesting the burden of hereditary cancers may be significant at the population level. Hereditary breast and ovarian cancer syndrome (HBOCS) is an inherited condition, typically caused by mutations in the BReast CAncer (BRCA1 and BRCA2) genes, that increases the risk of developing cancer, particularly early-onset breast and ovarian cancers.[5–9]

While HBOCS-related cancer risk estimates vary in the literature,[10–14] robust studies[13,15] highlight the greater lifetime cancer risk faced by female BRCA1 and BRCA2 mutation carriers (69–72% for breast; 17–44% for ovarian)[13] compared to the general population (12% for breast; 1–2% for ovarian).[10] The risk increases as the number of affected family members increases.[13,16] Male carriers face up to an 8% cumulative risk of breast cancer, whereas the general male population risk is much lower at 0.1%.[10,17] Prior research also suggests that these hereditary cancers exhibit more aggressive features (i.e., high-grade, advanced stage).[10] BRCA1 mutations have been associated with higher grade and triple-negative breast cancer, an aggressive subtype.[18,19] BRCA1 and BRCA2 mutations have also been linked to high-grade, serous/epithelial ovarian cancers.[20] Due to the elevated cancer risk associated with HBOCS, guidelines based on risk factors have been developed to identify affected individuals and recommend preventative measures and/or treatment.[4,21]

Existing studies have focused on BRCA1 and BRCA2 mutation carriers[11–13,15,22] and specific populations, such as women.[12,13,15] Given recent evidence of significant associations between HBOCS and other genes, as well as variants of unknown significance, particularly among racial/ethnic minorities,[14,23–26] the impact of HBOCS-related cancers may be underestimated. While the push toward multi-gene panel testing in hereditary cancer risk assessment[27,28] may help increase sensitivity toward detecting affected individuals, the lack of routine data collection of risk factors

[a]Rutgers Cancer Institute, New Brunswick, NJ [b]Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Piscataway, NJ, United States
Address correspondence to Caroline B. Morales, MPH, Department of Cancer Registry-Research, Rutgers Cancer Institute, 135 East State Street, Trenton NJ, 08625. Email: cbm72@cinj.rutgers.edu.

associated with hereditary cancers remains a barrier to assessing the true population burden. In one study, the Utah Cancer Registry measured the population-based prevalence of cancers meeting HBOCS-related criteria;[29] however, its results may not be applicable or reproducible in states with more diverse populations and no population-based system for collecting family history data, like New Jersey (NJ).

Despite anecdotal evidence of high genetic testing rates among NJ women (Lisa Paddock, MPH, PhD, personal communication, August 31, 2022), the incidence of HBOCS in NJ remains unknown due to the lack of available genetic and family history data. Without knowing the incidence, implementing cancer control and prevention efforts (i.e., screening, preventative surgery) targeting those at risk is more challenging. To fill this gap, we sought to estimate the incidence of HBOCS-related breast and ovarian cancers among NJ residents from 2015–2019 using risk factor information (e.g., primary site, age at diagnosis, breast and ovarian subtype) available at the New Jersey State Cancer Registry (NJSCR). To further characterize the burden, we compared the race/ethnicity and stage patterns of HBOCS-related breast and ovarian cancers and all breast and ovarian cancers.

Materials and Methods

Data Source

In this population-based cohort study, we utilized NJSCR data for New Jersey residents diagnosed with cancer between 2015–2019 to generate incidence statistics. Study variables were obtained from the NJSCR SEER*Stat Database analytic file created in 2023.

Study Population

The study sample (N=55,437) included all breast and ovarian cancers diagnosed among NJ women and men from 2015 through 2019. We included invasive breast, ovarian, fallopian tube, and peritoneal cancers as well as ductal carcinoma *in situ* breast cancers. All cancers with ICD-O-3 histology codes 9050–9055, 9140, and 9590–9992 were excluded per SEER standard.[30] Three sarcoma ovarian cancer cases and 1 case with unclear histology were excluded from all analyses.

Measures

We followed genetic testing criteria from the 2021 National Comprehensive Cancer Network's (NCCN) Breast, Ovarian, and Pancreatic Genetic/Familial High-Risk Assessment guidelines[21] to identify cancers with diagnostic characteristics of hereditary breast and ovarian cancers. Five groups based on these characteristics were created using a combination of variables as defined in Table 1. Cancers fitting any group criteria were flagged as possibly being HBOCS-related (yes/no) and referred to as "possible-HBOCS (pHBOCS)" throughout this paper.

From the NJSCR, we obtained site recode ICD-O-3/WHO2008, primary site [North American Association of Central Cancer Registries item (NAACCR #400)], histologic type ICD-O-3 (NAACCR #522), year of diagnosis (NAACCR #390), sex (NAACCR #220), age at diagnosis (NAACCR #230), sequence number—Central (NAACCR #380), breast subtype (2011+), race, Hispanic origin, stage for analysis, and county of patient residency at diagnosis variables. Hispanic origin is based on the NAACCR Hispanic Identification Algorithm (NHIA, NAACCR #191), which improves classification of ethnicity.[31] Sequence number-Central indicates the number and sequence of all primary tumors diagnosed within a person's lifetime. Breast subtype (2011+) includes four subtypes based on joint hormone receptor [HR; estrogen receptor (ER) and progesterone receptor (PR)] and human epidermal growth factor 2 (HER2) receptor statuses for breast cancer cases. Based on the SEER

Table 1. pHBOCS Group Variables and Definitions

pHBOCS Group	*SEER Variable(s)*	*Definition*
Early onset breast[a] cancer	Age at diagnosis	Breast cancer diagnosed ≤ age 45 (female only)
Multiple breast primaries	Sequence number—central (01–05)[b] and age at diagnosis	First breast cancer diagnosed between ages 46–50, with a subsequent breast cancer diagnosed at any age (female only)
Triple-negative breast cancer	Breast subtype (2011+)[c] and age at diagnosis	Estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and human epidermal growth factor receptor 2 (HER2)-negative and diagnosed ≤ age 60 (female only)
Male breast cancer	Sex	Male breast cancer diagnosed at any age
Epithelial ovarian cancer	Primary site and histologic type ICD-O-3	Epithelial ovarian cancer (or fallopian tube or peritoneal cancer)[d] at any age

Abbreviation: pHBOCS, possible Hereditary Breast and Ovarian Cancer Syndrome; SEER, Surveillance, Epidemiology, and End Results Program

[a] Site recode ICD-O-3/WHO2008 is 26000 for breast cancer.

[b] Sequence number-central indicates the sequence of all reportable cancers over a person's lifetime (01–05: more than 1 primary cancer).

[c] Breast subtype (2011+) is defined by joint hormone receptor (HR; ER and PR) and HER2 status. This variable was used to determine triple negative (yes/no/unknown or borderline) type status. Unknown included those with unavailable status or blanks.

[d] Defined as invasive ovarian, fallopian tube, and peritoneal cancer sites C569, C481–C482, C570 and epithelial ovarian ICD-O-3 histologies: 8001, 8010, 8020, 8021, 8041, 8050, 8070, 8071, 8075, 8120, 8140, 8246, 8260, 8310, 8320, 8323, 8380, 8381, 8440, 8441, 8442, 8460, 8461, 8462, 8470, 8472, 8474, 8480, 8490, 8574, 8950, 8980, 9000.

Table 2. Descriptive Statistics of Cancer Cases Meeting HBOCS Characteristics Among New Jersey Residents Overall and by Race and Ethnicity, 2015–2019, (N=55,437)

	Race and Ethnicity[a]				
HBOCS Characteristics by cancer	**Overall N (%)**	**Non-Hispanic White n (%)**	**Non-Hispanic Black n (%)**	**Non-Hispanic Asian and Pacific Islander n (%)**	**Hispanic n (%)**
Breast cancer,[b] total, n	51 448	35 515	6323	3636	5699
Age at diagnosis					
≤ 45 years	5889 (11.5)	3171 (9.0)	802 (12.8)	753 (20.8)	1113 (19.6)
> 45 years	45 157 (88.5)	32 041 (91.0)	5462 (87.2)	2866 (79.2)	4567 (80.4)
Sex					
Male	402 (0.8)	303 (0.9)	59 (0.9)	17 (0.5)	19 (0.3)
Female	51 046 (99.2)	35 212 (99.1)	6264 (99.1)	3619 (99.5)	5680 (99.7)
Multiple primaries[c]					
Yes	1534 (3.0)	1087 (3.1)	178 (2.8)	127 (3.5)	141 (2.5)
No	49 512 (97.0)	34 125 (96.9)	6086 (97.2)	3492 (96.5)	5,539 (97.5)
Triple-negative[d]					
Yes	1889 (3.7)	982 (2.8)	460 (7.3)	131 (3.6)	306 (5.4)
No	35 412 (69.4)	25 075 (71.2)	4082 (65.2)	2303 (63.6)	3812 (67.1)
Unknown/ Borderline	13 745 (26.9)	9155 (26.0)	1,722 (27.5)	1185 (32.7)	1562 (27.5)
Ovarian Cancer total, n	3989	2784	420	300	466
Epithelial ovarian,[e] fallopian tube, or primary peritoneal					
Yes	3588 (89.9)	2569 (92.3)	335 (79.8)	278 (92.7)	390 (83.7)
No	401 (10.1)	215 (7.7)	85 (20.2)	22 (7.3)	76 (16.3)
pHBOCS breast and ovarian cancers[f]					
pHBOCS breast only[g]	9090 (17.7)	5252 (14.8)	1345 (21.3)	975 (26.8)	1460 (25.6)
pHBOCS breast and ovarian	12 679 (22.9)	7821 (20.4)	1680 (24.9)	1253 (31.8)	1851 (30.0)
Non-pHBOCS	42 758 (77.1)	30 478 (79.6)	5063 (75.1)	2683 (68.2)	4314 (70.0)

Abbreviation: HBOCS, Hereditary Breast and Ovarian Cancer Syndrome; pHBOCS, possible HBOCS

[a] Race and ethnicity: Includes Hispanic of any race.

[b] Includes invasive and ductal carcinoma in situ.

[c] Multiple primaries defined dichotomously (yes/no) as first breast primary cancer diagnosed at age 46–50 with subsequent breast primary at any age (female only).

[d] Triple-negative breast cancer defined (yes/no/unknown or borderline) as cancers lacking estrogen (ER), progesterone (PR), and hormone receptors (HER2) and diagnosed ≤ age 60. Unknown included those with unavailable status or blanks (female only).

[e] Epithelial ovarian cancer includes invasive mucinous and non-mucinous histologies.

[f] Includes cancers meeting at least one of the following criteria: male breast cancer, breast cancer diagnosed ≤ age 45, first primary breast cancer diagnosed age 46–50 with 2+ primary breast cancer at any age, triple-negative and diagnosed ≤ age 60, or any epithelial ovarian cancer (including peritoneal and fallopian tube).

[g] Category represents the pHBOCS-related breast cancers out of all diagnosed breast cancers alone.

Note: Most rows may not sum to total as cases may meet more than one criterion; categories are not mutually exclusive.

software algorithm, if all three receptor statuses are negative, then the subtype is HER2-/HR- (triple-negative).[32] Race and ethnicity were combined and recoded as four mutually exclusive categories: non-Hispanic White, non-Hispanic Black, non-Hispanic Asian and Pacific Islander (API), and Hispanic (any race). The SEER stage for analysis variable (combination of NAACCR numbers 759, 760, 762, and 3020),[31] representing stage at diagnosis or treatment, was recategorized as *in situ* (for breast only), local, regional, distant, and unstaged. The county variable included all 21 NJ counties.

Statistical Analysis

We presented descriptive statistics of HBOCS-related characteristics for the NJ population by cancer type and race/ethnicity. We generated age-adjusted incidence rates (per 100,000 population) for pHBOCS breast and ovarian cancers, overall, by county and by race. County rates were

generated for each racial/ethnic group. RR by stage were analyzed to determine severity of disease within pHBOCS cancers. pHBOCS rate ratios (RR) by race and stage were compared to those of all cancers. We calculated combined male and female rates and female-only rates. Rates were age-adjusted to the 2000 US standard population (19 age groups- Census P25-1130), and 95% confidence intervals (CI) for rates and rate ratios were generated using the Tiwari et al 2006 estimation method. Rate ratios were considered statistically significant at the 0.05 level. All statistics were generated using SEER*Stat 8.4.

Results

Sample Characteristics

Overall, 51,448 breast and 3,989 ovarian cancers (including fallopian tube and peritoneal cancers) were diagnosed among NJ residents from 2015–2019 (Table 2). Most breast cancers were diagnosed in women (99.2%), 3.7% were triple-negative, and 3% fit the multiple primary criteria. Most ovarian, fallopian tube, and peritoneal cancers met the pHBOCS criteria (89.9%). Nearly 18% of all breast cancers and 23% of all breast and ovarian cancers (N=12,679) fit the pHBOCS definition.

We observed differences in the percentages of HBOCS diagnostic characteristics by race/ethnicity. The proportion of cancers diagnosed at or before age 45 was higher among APIs (20.7%), Hispanics (19.5%), and Blacks (12.8%) compared to Whites (9.0%). Similarly, the proportion of triple-negative breast cancers was higher among Blacks (7.3%), APIs (3.6%), and Hispanics (5.4%) than Whites (2.8%). The proportion of pHBOCS cancers was also higher among all racial minority groups compared to Whites. However, the distribution of cancers by sex and multiple primary criteria was similar regardless of race/ethnicity.

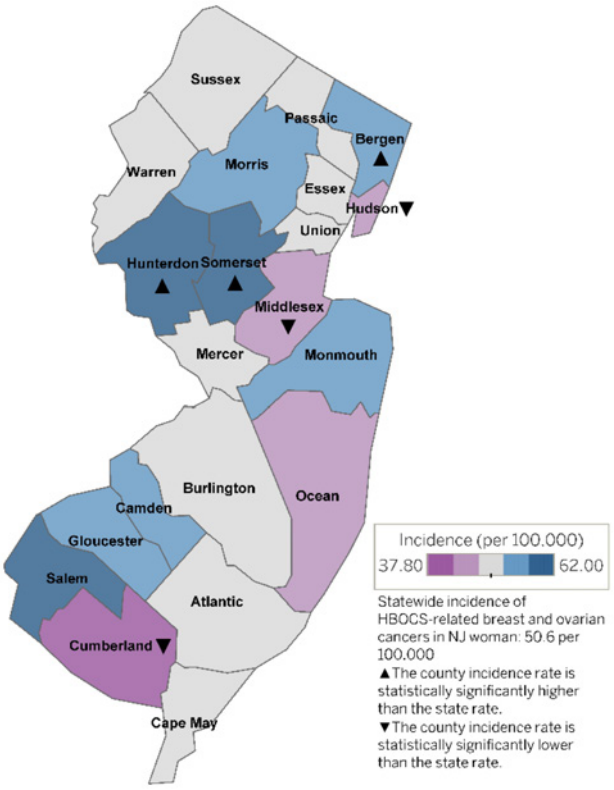
Age-adjusted Incidence by County

The age-adjusted incidence rate (AAIR) of pHBOCS breast and ovarian cancers among NJ women was 50.6 per 100,000 (95% CI: 49.6–51.5) from 2015–2019 (Figure 1). Relative to the statewide rate, Hunterdon (AAIR 62.0, 95% CI: 52.9–72.1), Somerset (AAIR 57.2, 95% CI: 52.0–62.7), and Bergen (AAIR 55.2, 95% CI: 52.2–58.4) counties had significantly higher rates, whereas Middlesex (AAIR 45.2, 95% CI: 42.3–48.2), Hudson (AAIR 43.0, 95% CI: 39.8–46.4), and Cumberland (AAIR 37.8, 95% CI: 31.5–44.9) counties had significantly lower rates.

Age-adjusted Incidence by Race and County

Among Hispanic women, the incidence of pHBOCS breast and ovarian cancers was significantly higher in Hunterdon County (AAIR 115, 95% CI: 68.4–181.1) but lower in Hudson County (AAIR 33.8, 95% CI: 29.6–38.4), relative to the statewide Hispanic rate (AAIR 40.6, 95% CI: 38.7–42.6) (not shown). The rate was significantly lower in Ocean County for Whites (AAIR 45.3, 95% CI: 41.3–49.7), Cumberland County for Blacks (AAIR 25.5, 95% CI: 14.6–41.3), and Atlantic County for APIs (AAIR 26.3, 95% CI: 14.8–43.5) when compared to the statewide rate of their respective racial/ethnic group.

Figure 1. Age-adjusted Cancer Incidence (per 100,000) of Possible HBOCS-related Female Breast and Ovarian Cancers Diagnosed in NJ by County, 2015–2019



All Cancers versus pHBOCS-related Cancers

Racial/ethnic patterns were similar for all female breast and ovarian cancers and pHBOCS female breast and ovarian cancers (not shown). Among all female breast and ovarian cancers, the AAIRs were significantly lower for Blacks (RR=0.90, 95% CI:0.87–0.92), APIs (RR=0.77, 95% CI:0.75–0.80), and Hispanics (RR=0.74, 95% CI:0.72–0.76) relative to Whites. Likewise, among the pHBOCS cancers, the AAIRs were significantly lower for Black (RR=0.90, 95% CI:0.85–0.95), API (RR=0.85, 95% CI:0.80–0.91), and Hispanic women (RR=0.74, 95% CI:0.70–0.79) compared to White women.

Table 3 compares AAIRs and RRs by stage for female-only breast cancers. Similar to all breast cancers (RR=2.18, 95% CI:2.14–2.23), the incidence of local-stage cancers among the pHBOCS breast cancers was significantly higher than the incidence of *in situ* cancers (RR=1.86, 95% CI:1.76–1.98). The incidence of distant-stage breast cancers was significantly lower (All: RR=0.22, 95% CI:0.21–0.23 versus pHBOCS: RR=0.19, 95% CI:0.17–0.21) than *in situ*. The rate of regional-stage cancers was significantly higher compared to the rate of *in situ* cancers among the pHBOCS breast cancers (RR=1.16, 95% CI:1.09–1.24) but significantly lower among all breast cancers (RR=0.84, 95% CI:0.82–0.87). With respect to ovarian cancers (Table 4), the incidence of distant-stage cancers was significantly higher than the incidence of local-stage cancers among all ovarian cancers (RR=2.59, 95% CI:2.37–2.83) and the pHBOCS ovarian cancers (RR=3.22,

Table 3. Female Age-adjusted Cancer Incidence and Rate Ratios by Stage: All Breast Cancers vs. pHBOCS[a] Breast Cancers, 2015–2019 (N=51 045)

	All Breast Cancers					*pHBOCS Breast Cancers*				
Stage[b]	**Count**	**Age-adjusted[c] incidence (per 100,000)**	**(95% CI)[d]**	**Rate* ratio**	**(95% CI)**	**Count**	**Age-adjusted incidence (per 100 000)**	**(95% CI)**	**Rate ratio**	**(95% CI)**
In Situ	11 411	41.5	(40.7–42.3)	Ref		1725	8.9	(8.5–9.3)	Ref	
Local	26 065	90.6	(89.5–91.8)	2.18	(2.14–2.23)	3438	16.6	(16.0–17.2)	1.86	(1.76–1.98)
Regional	9549	35.0	(34.3–35.8)	0.84	(0.82–0.87)	2072	10.4	(9.9–10.8)	1.16	(1.09–1.24)
Distant	2644	9.0	(8.6–9.3)	0.22	(0.21–0.23)	354	1.7	(1.5–1.9)	0.19	(0.17–0.21)
Unstaged	1376	4.6	(4.3–4.8)	0.11	(0.10–0.12)	170	0.8	(0.7–1.0)	0.09	(0.08–0.11)

Abbreviation: CI, Confidence interval; pHBOCS, possible Hereditary Breast and Ovarian Cancer Syndrome

[a] pHBOCS includes cancers meeting at least one of the following criteria: breast cancer diagnosed ≤ age 45, first primary breast cancer age 46–50 with 2+ primary breast cancer at any age, triple-negative diagnosed ≤ age 60, or any epithelial ovarian cancer (including peritoneal and fallopian tube).

[b] The SEER stage for analysis variable uses an algorithm to combine multiple stage variables (combination of NAACCR item #759, 760, 762, and 3020), representing stage at diagnosis or treatment.

[c] Age-adjusted to 2000 U.S. standard population.

[d] 95% CI generated by Tiwari et al., 2006 estimation method for age-adjusted rates.

*Significant at p<.05.

Table 4. Age-adjusted Cancer Incidence and Rate Ratios by Stage: All Ovarian Cancers vs. pHBOCS[a] Ovarian Cancers, 2015–2019, (N=3965)

	All Ovarian Cancers					*pHBOCS Ovarian Cancers*				
Stage[b]	**Count**	**Age-adjusted[c] incidence (per 100,000)**	**(95% CI)[d]**	**Rate* ratio**	**(95% CI)**	**Count**	**Age-adjusted incidence (per 100 000)**	**(95% CI)**	**Rate ratio**	**(95% CI)**
Local	718	2.7	(2.5–2.9)	Ref		573	2.1	(1.9–2.3)	Ref	
Regional	791	2.8	(2.6–3.0)	1.01	(0.91–1.12)	736	2.5	(2.4–2.7)	1.21	(1.08–1.36)
Distant	2115	7.1	(6.8–7.4)	2.59	(2.37–2.83)	2034	6.8	(6.5–7.1)	3.22	(2.93–3.55)
Unstaged	341	1.1	(1.0–1.3)	0.41	(0.36–0.47)	222	0.7	(0.6–0.8)	0.35	(0.30–0.41)

Abbreviation: CI, Confidence interval; pHBOCS, possible Hereditary Breast and Ovarian Cancer Syndrome

[a] pHBOCS includes cancers meeting at least one of the following criteria: breast cancer diagnosed ≤ age 45, first primary breast cancer age 46–50 with 2+ primary breast cancer at any age, triple-negative diagnosed ≤ age 60, or any epithelial ovarian cancer (including peritoneal and fallopian tube).

[b] The SEER stage for analysis variable uses an algorithm to combine multiple stage variables (combination of NAACCR item #759, 760, 762, and 3020), representing stage at diagnosis or treatment.

[c] Age-adjusted to 2000 U.S. standard population.

[d] 95% CI generated by Tiwari et al., 2006 estimation method for age-adjusted rates.

*Significant at p<.05.

95% CI:2.93–3.55). The rate of regional-stage ovarian cancers was significantly higher than local for pHBOCS cancers (RR=1.21, 95% CI:1.08–1.36).

Discussion

We demonstrated a novel way of estimating HBOCS cancer incidence in NJ using cancer registry data and further characterized the HBOCS burden by evaluating racial/ethnic and stage distributions. Racial/ethnic minorities had a disproportionate burden of early-onset and triple-negative breast cancers compared to their White counterparts. They also had larger proportions of breast and ovarian cancers meeting pHBOCS criteria. However, the incidence of pHBOCS breast and ovarian cancers was significantly lower among API, Hispanic, and Black women compared to White women. Racial/ethnic patterns observed among the pHBOCS cancers mirrored those of all breast and ovarian cancers. In analyses that included male breast cancers, rates were dramatically reduced due to the low incidence of pHBOCS cancers among men, but rate ratios by race/ethnicity reflected similar racial differences in rates. The incidence of late-stage cancers, particularly ovarian, was more pronounced among the pHBOCS cancers than all

cancers. Therefore, the burden of pHBOCS cancers among NJ women may be substantive given the more pronounced rates of advanced stage cancers. Lastly, we observed racial and county-level differences in rates and that pHBOCS county rates were modified by race.

With respect to racial/ethnic differences in the distribution of HBOCS characteristics, our findings are consistent with the literature. Hendrick et al[33] found a higher proportion of breast cancers diagnosed under age 50 among Black, Hispanic, and API women relative to non-Hispanic White women, though their study was limited to invasive cancers. Population-based studies have also shown that the highest incidence of triple-negative breast cancer occurs among Black women relative to any other racial group.[34,35] Another study[36] found higher odds of triple-negative breast cancer among Hispanic women relative to White women, but no difference remained after controlling for age and stage. Moreover, Greenberg et al's study[29] found similar racial/ethnic distributions of HBOCS diagnostic risk factors in a US sample of breast and ovarian cancers diagnosed between 2010–2015. They also reported higher proportions of breast cancers meeting any criteria among racial minority groups. Despite observing higher proportions of multiple HBOCS risk factors among minority women, their pHBOCS cancer rates were lower than the White pHBOCS rate. Higher incidence of omitted risk factors like Ashkenazi Jewish ancestry among non-Hispanic Whites and the fact that we only used diagnostic criteria to classify pHBOCS may partially explain why we observed the greatest pHBOCS incidence among White cancers. Another explanation for the discrepancy may be that there are more minority cases with unknown race/ethnicity, making counts for minority group cases lower and their proportions with HBOCS characteristics higher. Likewise, if White women had fewer unknowns, their case counts would be higher and the proportion with HBOCS characteristics would be lower. Studies have noted that misclassification of race/ethnicity in cancer registry data is more likely among non-White race/ethnicities[37] and that missing data is more prevalent in the cancer records of racial/ethnic minorities.[38]

Our stage analysis for pHBOCS cancers aligns with previous studies associating BRCA-related HBOCS with high-grade, serous/epithelial ovarian cancers, which are typically diagnosed at advanced stages.[10,20] We observed a more pronounced pattern of late-stage cancers among women with ovarian cancers meeting HBOCS diagnostic criteria, yet our findings are not limited to mutations in 1 gene. However, breast cancer patterns by stage reflect what is observed for breast cancers overall and ovarian cancers are known to be diagnosed at late-stage, so higher late-stage rates compared to early-stage are expected.

Our study had several limitations. As with all studies utilizing cancer registry data, misclassification by race, ethnicity, and county at diagnosis is possible. However, our race, ethnicity, and county data meet the NAACCR quality standards for gold certification and follow the NAACCR race/ethnicity coding algorithm, reducing the potential for misclassification bias. Nevertheless, our racial comparisons are limited to 4 groups, as we excluded American Indian, Alaska Native, and other racial groups with small sample sizes to avoid unstable rates. Our choice of statistical software limited our pHBOCS definition as it limited our ability to define age categories for pHBOCS groups. Because SEER*Stat uses 5-year age groups for calculating rates (i.e., 40–44, 45–49, … 55–59, 60–65, etc.), our pHBOCS rates exclude breast cancers diagnosed among women at age 45 and triple-negative breast cancers diagnosed among women at age 60. Consequently, some cancers meeting the HBOCS risk assessment criteria may have been misclassified, underestimating the rate of pHBOCS cancers. Likewise, this software precluded comparisons between pHBOCS cancers and non-pHBOCS cancers. We compared patterns of pHBOCS cancers to all cancers instead, so our race/ethnicity and stage analyses may be biased. Because we did not include all the cancers associated with HBOCS, our estimate may not reflect the true burden. Recent NCCN guidelines, released after the start of our study, recommend less restrictive criteria, such as genetic testing for individuals diagnosed with breast cancer at or below age 50, multiple breast primaries regardless of age, and hormone receptor positive breast cancers receiving certain treatments.[39] However, including hormone receptor positive breast cancers may produce estimates that incorrectly classify low-risk individuals as having HBOCS. Furthermore, state privacy laws are a significant limitation to genetic risk factor surveillance. For this reason, we could not account for genetic risk factors in our definition of HBOCS-related cancers. Cancers identified as pHBOCS in this study may come from individuals lacking high-risk genes, so the real HBOCS cancer rate is likely lower. We also could not account for family history, as reporting is not required by central cancer registries. Because registries do not uniformly collect lifestyle/environmental risk factor information associated with non-hereditary breast and ovarian cancers (i.e., obesity, smoking, parity, etc.),[40–42] we could not adjust for confounders, so pHBOCS proportions (particularly ovarian) and rates are overestimated.

The present study is strengthened by its use of high-quality incidence data from a population-based cancer registry, minimizing potential sampling bias. While the lack of genetic and family history information precludes an accurate estimate of the true burden, our focus on diagnostic characteristics may be informative when assessing racial/ethnic differences. Relying on results from genetic panels alone may result in misclassification of high-risk individuals, particularly those from minority populations.[43] Likewise, we included men in our sample, a group that is underrepresented in studies related to this syndrome.

Conclusions

Our findings represent an initial estimate of the cancer burden associated with HBOCS in NJ. Quantifying the impact of hereditary cancers and documenting the challenges in doing so represent the first steps toward identifying the true population burden. Identifying areas with greater incidence can inform future policies that address resource allocation for prevention efforts. Our study showed potential racial and county-level differences in HBOCS-related

cancer incidence and a possible greater burden of late-stage cancers among those meeting HBOCS genetic testing criteria. Given the limitations of our study, findings must be confirmed by further research. Due to privacy laws, researchers may seek to link cancer registry data to genetic testing companies in an anonymized fashion (i.e., encrypted IDs) to better classify HBOCS-related cancers.[44] Survey data from selected participants can be used to obtain family history information. Future studies may also expand to include other HBOCS-related cancers (i.e., prostate and pancreatic) or explore the impact of changing risk assessment guidelines.

Acknowledgements

We would like to thank Heather Stabinsky, MSEd, CTR, Karen Pawlish, Sc.D, MPH, and Jie Li, MPH at the NJSCR.

This project was supported by the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services (HHS) as part of a financial assistance award funded by CDC/HHS and by non-government source(s). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, CDC/HHS, or the US government. The NJSCR is funded by the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (#75N91021D00009) and CDC's National Program of Cancer Registries (#5NU58DP006279) with additional support from the State of New Jersey and the Rutgers Cancer Institute.

NCCN makes no warranties of any kind whatsoever regarding its content use or application and disclaims any responsibility for its application or use thereof.

References:

1. U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on 2024 submission data (1999–2023): United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. June 2025. **https://www.cdc.gov/cancer/dataviz**
2. American Cancer Society. Cancer Facts & Figures 2024. Atlanta: American Cancer Society; 2024.
3. Ramus SJ, Gayther SA. The contribution of BRCA1 and BRCA2 to ovarian cancer. *Mol Oncol.* 2009;3(2):138–50. doi:10.1016/j.molonc.2009.02.001
4. Owens DK, Davidson KW, Krist AH, et al. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* 2019;322(7):652–665. doi:10.1001/jama.2019.10987
5. National Institutes of Health, National Cancer Institute. *NCI Dictionary of Cancer Terms.* Accessed August 20, 2022. **https://www.cancer.gov/publications/dictionaries/cancer-terms/**
6. Beebe-Dimmer JL, Kapron AL, Fraser AM, Smith KR, Cooney KA. Risk of prostate cancer associated with familial and hereditary cancer syndromes. *J Clin Oncol* 2020;38(16):1807–1813. doi:10.1200/jco.19.02808
7. Ibrahim M, Yadav S, Ogunleye F, Zakalik D. Male BRCA mutation carriers: clinical characteristics and cancer spectrum. *BMC Cancer* 2018;18(1):179. doi:10.1186/s12885-018-4098-y
8. McGarry JL, Creavin B, Kelly ME, Gallagher TK. Risk of pancreatic ductal adenocarcinoma associated with carriage of BRCA1 and/or BRCA2 mutation: A systematic review and meta-analysis. *J Surg Oncol* 2022; 126(6):1028–1037. doi:10.1002/jso.26994
9. Pilarski R. The role of BRCA testing in hereditary pancreatic and prostate cancer families. *Am Soc Clin Oncol Educ Book* 2019;39:79-86. doi:10.1200/edbk_238977
10. Petrucelli N, Daly MB, Pal T. BRCA1- and BRCA2-associated hereditary breast and ovarian cancer. In: Adam MP, Mirzaa GM, Pagon RA, et al, editors. *GeneReviews*. University of Washington, Seattle; 1993–2025; 1998 Sep 4. Accessed August 20, 2025. **https://www.ncbi.nlm.nih.gov/books/NBK1247/**
11. Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: A combined analysis of 22 studies. *Am J Hum Genet.* 2003;72(5):1117–30. doi:10.1086/375033
12. Mavaddat N, Peock S, Frost D, et al. Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. *J Natl Cancer Inst* 2013;105(11):812–22. doi:10.1093/jnci/djt095
13. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* 2017;317(23):2402–2416. doi:10.1001/jama.2017.7112
14. Yoshida R. Hereditary breast and ovarian cancer (HBOC): Review of its molecular characteristics, screening, treatment, and prognosis. *Breast Cancer* 2021;28(6):1167–1180. doi:10.1007/s12282-020-01148-2
15. Chen J, Bae E, Zhang L, et al. Penetrance of breast and ovarian cancer in women who carry a BRCA1/2 mutation and do not use risk-reducing salpingo-oophorectomy: An updated meta-analysis. *JNCI Cancer Spectr* 2020;4(4):pkaa029. doi:10.1093/jncics/pkaa029
16. Zheng G, Yu H, Kanerva A, Försti A, Sundquist K, Hemminki K. Familial risks of ovarian cancer by age at diagnosis, proband type and histology. *PLoS One* 2018;13(10):e0205000. doi:10.1371/journal.pone.0205000
17. Evans DG, Susnerwala I, Dawson J, Woodward E, Maher ER, Lalloo F. Risk of breast cancer in male BRCA2 carriers. *J Med Genet* 2010;47(10):710–711. doi:10.1136/jmg.2009.075176
18. Mavaddat N, Barrowdale D, Andrulis IL, et al. Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: Results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiol Biomarkers Prev* 2012;21(1):134–47. doi:10.1158/1055-9965.Epi-11-0775
19. Lee E, McKean-Cowdin R, Ma H, et al. Characteristics of triple-negative breast cancer in patients with a BRCA1 mutation: Results from a population-based study of young women. *J Clin Oncol* 2011;29(33):4373–80. doi:10.1200/jco.2010.33.6446
20. McLaughlin JR, Rosen B, Moody J, et al. Long-term ovarian cancer survival associated with mutation in BRCA1 or BRCA2. *J Natl Cancer Inst* 2013;105(2):141–148. doi:10.1093/jnci/djs494
21. Daly MB, Pal T, Berry MP, et al. Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2021;19(1):77–102. doi:10.6004/jnccn.2021.0001
22. Chen S, Iversen ES, Friebel T, et al. Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *J Clin Oncol* 2006;24(6):863–71. doi:10.1200/jco.2005.03.6772
23. Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med* 2014;371(6):497–506. doi:10.1056/NEJMoa1400382
24. Tatineni S, Tarockoff M, Abdallah N, et al. Racial and ethnic variation in multigene panel testing in a cohort of BRCA1/2-negative individuals who had genetic testing in a large urban comprehensive cancer center. *Cancer Med* 2022;11(6):1465–1473. doi:10.1002/cam4.4541
25. Caswell-Jin JL, Gupta T, Hall E, et al. Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. *Genet Med* 2018;20(2):234–239. doi:10.1038/gim.2017.96
26. da Costa E Silva Carvalho S, Cury NM, Brotto DB, et al. Germline variants in DNA repair genes associated with hereditary breast and ovarian cancer syndrome: analysis of a 21 gene panel in the Brazilian population. *BMC Med Genomics* 2020;13(1):21. doi:10.1186/s12920-019-0652-y
27. Yoo J, Lee GD, Kim JH, et al. Clinical validity of next-generation sequencing multi-gene panel testing for detecting pathogenic variants in patients with hereditary breast-ovarian cancer syndrome. *Ann Lab Med* 2020;40(2):148–154. doi:10.3343/alm.2020.40.2.148

28. Desmond A, Kurian AW, Gabree M, et al. Clinical actionability of multigene panel testing for hereditary breast and ovarian cancer risk assessment. *JAMA Oncol* 2015;1(7):943–951. doi:10.1001/jamaoncol.2015.2690
29. Greenberg S, Buys SS, Edwards SL, et al. Population prevalence of individuals meeting criteria for hereditary breast and ovarian cancer testing. *Cancer Med* 2019;8(15):6789–6798. doi:10.1002/cam4.2534
30. National Cancer Institute's Surveillance, Epidemiology, and End Results Program. Site Recode ICD-O-3/WHO 2008 Definition. Accessed January 12, 2025. **https://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html#footnotes**
31. Thornton ML, (ed). Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Version 24, 25th ed. Springfield, Ill: North American Association of Central Cancer Registries, July 2023. Rev. July 2023, August 2023, October 2023, February 2024, March 2024. Available from: **https://apps.naaccr.org/data-dictionary/**
32. National Cancer Institute's Surveillance, Epidemiology, and End Results Program. Breast Subtype (2010+) - SEER Documentation. Accessed February 4, 2024. **https://seer.cancer.gov/seerstat/databases/ssf/breast-subtype.html**
33. Hendrick RE, Monticciolo DL, Biggs KW, Malak SF. Age distributions of breast cancer diagnosis and mortality by race and ethnicity in US women. *Cancer* 2021;127(23):4384–4392. doi:10.1002/cncr.33846
34. Zhang W, Bai Y, Sun C, Lv Z, Wang S. Racial and regional disparities of triple negative breast cancer incidence rates in the United States: An analysis of 2011-2019 NPCR and SEER incidence data. *Front Public Health* 2022;10:1058722. doi:10.3389/fpubh.2022.1058722
35. Du XL, Li Z. Incidence trends in triple-negative breast cancer among women in the United States from 2010 to 2019 by race/ethnicity, age and tumor stage. *Am J Cancer Res* 2023;13(2):678–691.
36. Scott LC, Mobley LR, Kuo TM, Il'yasova D. Update on triple-negative breast cancer disparities for the United States: A population-based study from the United States Cancer Statistics database, 2010 through 2014. *Cancer* 2019;125(19):3412–3417. doi:10.1002/cncr.32207
37. Gomez SL, Glaser SL. Misclassification of race/ethnicity in a population-based cancer registry (United States). *Cancer Causes Control* 2006;17(6):771–781. doi:10.1007/s10552-006-0013-y
38. Yang DX, Khera R, Miccio JA, et al. Prevalence of missing data in the National Cancer Database and association with overall survival. *JAMA Netw Open* 2021;4(3):e211793. doi:10.1001/jamanetworkopen.2021.1793
39. National Comprehensive Cancer Network Inc. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) for Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic V.3.2023. Aug 8, 2023. **https://www.nccn.org/**
40. Leitzmann MF, Koebnick C, Danforth KN, et al. Body mass index and risk of ovarian cancer. *Cancer* 2009;115(4):812–822. doi:10.1002/cncr.24086
41. Ali AT, Al-Ani O, Al-Ani F. Epidemiology and risk factors for ovarian cancer. *Prz Menopauzalny* 2023;22(2):93–104. doi:10.5114/pm.2023.128661
42. American Cancer Society. Lifestyle-related breast cancer risk factors. Accessed January 9, 2026. **https://www.cancer.org/cancer/types/breast-cancer/risk-and-prevention/lifestyle-related-breast-cancer-risk-factors.html**
43. Liu Q, Yao S, Zhao H, et al. Early-onset triple-negative breast cancer in multiracial/ethnic populations: Distinct trends of prevalence of truncation mutations. *Cancer Med* 2019;8(4):1845–1853. doi:10.1002/cam4.2047
44. Kurian AW, Abrahamse P, Furgal A, et al. Germline genetic testing after cancer diagnosis. *JAMA* 2023;330(1):43–51. doi:10.1001/jama.2023.9526

Original Article

Trends in Early-Onset and Late-Onset Colorectal Cancer by Subtype in New York State, 1976–2022

Rakkoo Chung[a];Margaret Gates Kuliszewski[a,b];Tabassum Insaf[a,b]

Abstract: Background: Despite the steadily decreasing incidence rate of colorectal cancer (CRC) overall, many studies have reported increasing incidence of CRC among younger people (age < 50), early-onset colorectal cancer (EOCRC), in the United States. However, fewer studies have examined the trends by CRC subtype. We sought to better understand the trends of CRC incidence by subtype and age at diagnosis (particularly before age 50 versus age 50 or later) in recent decades in New York State and in other areas in the United States. Methods: Using the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) data, we estimated incidence rates of CRC for 11 subtypes, for different age groups, and by stage (local versus regional and distant) from 1976 through 2022 in New York State, compared with the same estimates from 2000 through 2021 in all 22 United States SEER cancer registries. All incidence rates are per 100,000 and age-adjusted to the 2000 United States standard population. Results: While the overall incidence rate of all CRC combined decreased for ages 50 and above from 229.5 in 1976–1980 to 106.3 in 2021–2022, it increased for ages under 50 from 5.9 to 10.0 during the same period. For ages under 50, the subtype-specific incidence rates increased for appendix (from 0.04 to 1.3), rectum (from 1.3 to 3.1), and sigmoid colon (from 1.2 to 2.1), while the other subtypes showed relatively low and steady incidence rates. For ages 50 and above, most subtypes showed generally decreasing incidence rates, with the exception of appendix, which steadily increased from 0.3 to 3.8. Data from the 22 registries combined showed almost identical trends. Conclusion: While aggregate statistics show increasing incidence rates of EOCRC, it is important to consider trends by CRC subtype. We observed that the increasing trends of EOCRC are primarily driven by appendix, rectum, and sigmoid colon cancers, and that the incidence rate of appendix cancer has been increasing across all age groups. Future studies should consider differences by subtype of CRC, as this information may help elucidate factors contributing to recent trends in EOCRC incidence.

Key words: *Colorectal cancer; early-onset colorectal cancer; Incidence; Surveillance, Epidemiology, and End Results (SEER) Program*

Introduction

Colorectal cancer (CRC) was the fourth most common cancer (following female breast cancer, male prostate cancer, and lung and bronchus cancer) in New York State (NYS) in 2018–2022,[1] and nationally in 2025.[2] Many studies of CRC have reported that the incidence rate of early-onset colorectal cancer (EOCRC; CRC diagnosed before age 50) has been on the rise, even in countries such as the United States, where the overall incidence rate of CRC has been decreasing.[3]

Prior studies have investigated risk factors for EOCRC compared to those of late-onset colorectal cancer (LOCRC; CRC diagnosed at age 50 or later),[4] but most of these failed to examine which subtypes were driving the rising incidence of EOCRC. Only a few studies have highlighted the need for research of EOCRC by subtype. Chen and coauthors[5] showed that Singapore had witnessed different trajectories of colon cancer and rectal cancer from 1968 to 2019. Patel and coauthors[4] reviewed six studies of CRC in the United States and explored the incidence rates by location (i.e., right colon, left colon, and rectal) and discussed some risk factors for specific locations of tumors. Meyers and coauthors[6] reported that CRC incidence rates varied greatly by subtype (i.e., anus, appendix, distal colon, proximal colon, and rectum) and by age group in Australia from 1990 to 2020, and further explored the histological patterns by the subtype of CRC. Montminy and coauthors[7] demonstrated that appendix cancer significantly contributed to the increasing incidence rate of EOCRC in the United States and stressed the importance of stratifying by CRC subtype and histological subtype when analyzing incidence rates of EOCRC. Additionally, a study by Sung and coauthors noted that overestimation of colon cancer incidence probably occurred because appendiceal cancer was not separable from colon cancer in their data, and suggested that "detailed and systematic analyses of subtype-specific trends are warranted to elucidate reasons behind the concerning trends."[3] Therefore, in this analysis, we expanded on the previous studies and examined CRC incidence trends by age group, stage, and detailed subtype for NYS and 22 SEER registries.

[a]New York State Cancer Registry, New York State Department of Health, Albany, NY [b]Department of Epidemiology and Biostatistics, University at Albany, Albany, NY

Address correspondence to Rakkoo Chung, PhD, New York State Cancer Registry, New York State Department of Health, 150 Broadway, Suite 361, Albany, NY 12204. Email: rakkoo.chung@health.ny.gov.

This work was supported in part by cooperative agreement NU58DP007218 awarded to the New York State Department of Health by the Centers for Disease Control and Prevention and by Contract HHSN261201800005I, task order HHSN26100001, from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services. The opinions, findings, and conclusions expressed here are those of the authors and do not necessarily represent the official views of the funding agencies.

This study aimed to (1) explore trends of CRC incidence rates by subtype and age group from 1976 through 2022 in NYS Cancer Registry data and from 2000 through 2021 in data from 22 United States cancer registries,[8] and (2) highlight the importance of incorporating subtype-specific analyses when studying EOCRC.

Methods

Using the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) data from NYS, we examined incidence rates and average percent change (APC) in incidence rates for 11 subtypes[9] of malignant CRC (excluding unknown ages) and two stage categories (local versus regional and distant) from 1976 through 2022 in NYS (Table S1). We included all ages broadly categorized as EOCRC (ages 0–49) and LOCRC (ages 50 and above). For the APC analyses, we additionally examined age in five-year categories, except for the youngest individuals (ages 0–29), due to the small number of cases. Two age groups (0–29 and 30–44) are included for further comparison within the EOCRC cases. Because of the small numbers of cases for some specific subtypes, we estimated the incidence rates aggregated for five-year periods from 1976–1980 to 2016–2020 and the two-year period 2021–2022. The last period is excluded from some analyses when comparing the case counts. We also compared and contrasted the same estimates from 2000–2021 in 22 cancer registries (including the NYS Cancer Registry) in the United States. We used SEER*Stat 9.0.41[10] for all analyses, with incidence rates calculated per 100,000 persons and age-adjusted to the United States 2000 standard population.

Table S1. Subtypes of Colorectal Cancer and ICD-O-3 Site

Broad Subtype	*Subtype*	*ICD-O-3 Site*	*ICD-O-3 Histology*
Colon excluding rectum	Cecum	C180	Excluding 9050–9055, 9140, 9590–9993
	Appendix	C181	
	Ascending colon	C182	
	Hepatic flexure	C183	
	Transverse colon	C184	
	Splenic flexure	C185	
	Descending colon	C186	
	Sigmoid colon	C187	
	Large intestine, NOS	C188–C189, C260	
Rectum and rectosigmoid junction	Rectosigmoid junction	C199	
	Rectum	C209	

Source: **http://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html**. Accessed Jul 2025.

Results

Table S2 shows the number and percent of CRC cases diagnosed in NYS between 1976 and 2022 by age group, subtype, and year of diagnosis. Most CRC cases occurred in individuals aged 50 and above; however, the percentage of LOCRC cases decreased over time from 94.9% of all CRC cases diagnosed in NYS in 1976–1980 to 86.9% in 2021–2022 (Figure S1). The overall trend of CRC incidence (results not shown) is therefore similar to that for LOCRC, because the older population accounts for most CRC cases.

Examining trends in incidence rates by age category, Figure 1 displays the increasing trend in the overall incidence rate of EOCRC and the decreasing trend in the overall incidence rate of LOCRC. While the overall incidence rate of all subtypes combined decreased for LOCRC from 229.5 in 1976–1980 to 106.3 in 2021–2022, it increased for EOCRC from 5.9 to 10.0 during the same period. Similar trends were observed for both colon cancers and cancers in the rectum and rectosigmoid junction (rectal cancer, hereafter). Notably, rectal cancer accounts for a larger proportion of EOCRC than LOCRC, and the proportion of rectal cancer to total CRC has increased over time for both age groups. These patterns were also observed for the 22 SEER registries combined (Figure S2).

In analyses of 11 CRC subtypes, we observed diverse trends in the incidence rates of the subtypes for EOCRC and LOCRC (Figure 2). For EOCRC, the subtype-specific incidence rates clearly increased for only three subtypes: appendix (from 0.04 to 1.3), rectum (from 1.3 to 3.1), and sigmoid colon (from 1.2 to 2.1). In terms of the percent of diagnoses, these three subtypes increased from 42% of all EOCRC cases in 1976–1980 to 64% in 2016–2020, accounting for over 90% of the total increase in the EOCRC cases in the last four decades (results not shown). Incidence rates of the other subtypes remained low and were relatively steady or increased only slightly during the study period. Such trends differed from those for LOCRC, for which incidence rate trends for all subtypes except appendix were uniformly downward after the early 2000s, despite their diverse trends prior to that time. The exception was appendix, which increased from 0.3 to 3.8 over time throughout the years examined, although the incidence remained low. Aggregate data from the 22 cancer registries again showed similar trends (Figure S3).

Table S2. Characteristics of Colorectal Cancer (CRC) Cases by CRC Subtype and Year at Diagnosis, New York State, 1976–2022

	All cases	*Age <50*	*Age 50+*
Total, n (%)	**495 420 (100%)**	**37 137 (100%)**	**458 283 (100%)**
CRC subtype			
Colon excluding rectum	356 572 (72.0%)	23 657 (63.7%)	332 915 (72.6%)
Cecum	61 664 (12.4%)	3073 (8.3%)	58 591 (12.8%)
Appendix	7006 (1.4%)	2416 (6.5%)	4590 (1.0%)
Ascending colon	60 136 (12.1%)	2850 (7.7%)	57 286 (12.5%)
Hepatic flexure	12 794 (2.6%)	711 (1.9%)	12 083 (2.6%)
Transverse colon	29 536 (6.0%)	1794 (4.8%)	27 742 (6.1%)
Splenic flexure	9963 (2.0%)	801 (2.2%)	9162 (2.0%)
Descending colon	22 430 (4.5%)	1806 (4.9%)	20 624 (4.5%)
Sigmoid colon	98 506 (19.9%)	7516 (20.2%)	90 990 (19.9%)
Large intestine, NOS	54 537 (11.0%)	2690 (7.2%)	51 847 (11.3%)
Rectum & rectosigmoid junction	138 848 (28.0%)	13 480 (36.3%)	125 368 (27.4%)
Rectosigmoid junction	46 089 (9.3%)	3761 (10.1%)	42 328 (9.2%)
Rectum	92 759 (18.7%)	9719 (26.2%)	83 040 (18.1%)
Year of diagnosis			
1976 ~ 1980	56 166 (11.3%)	2 860 (7.7%)	53 306 (11.6%)
1981 ~ 1985	59 013 (11.9%)	2816 (7.6%)	56 197 (12.3%)
1986 ~ 1990	56 069 (11.3%)	2856 (7.7%)	53 213 (11.6%)
1991 ~ 1995	54 806 (11.1%)	3270 (8.8%)	51 536 (11.2%)
1996 ~ 2000	57 626 (11.6%)	3842 (10.3%)	53 784 (11.7%)
2001 ~ 2005	55 027 (11.1%)	4355 (11.7%)	50 672 (11.1%)
2006 ~ 2010	49 181 (9.9%)	4690 (12.6%)	44 491 (9.7%)
2011 ~ 2015	45 207 (9.1%)	4889 (13.2%)	40 318 (8.8%)
2016 ~ 2020*	44 219 (8.9%)	5186 (14.0%)	39 033 (8.5%)
2021 ~ 2022**	18 106 (3.7%)	2373 (6.4%)	15 733 (3.4%)

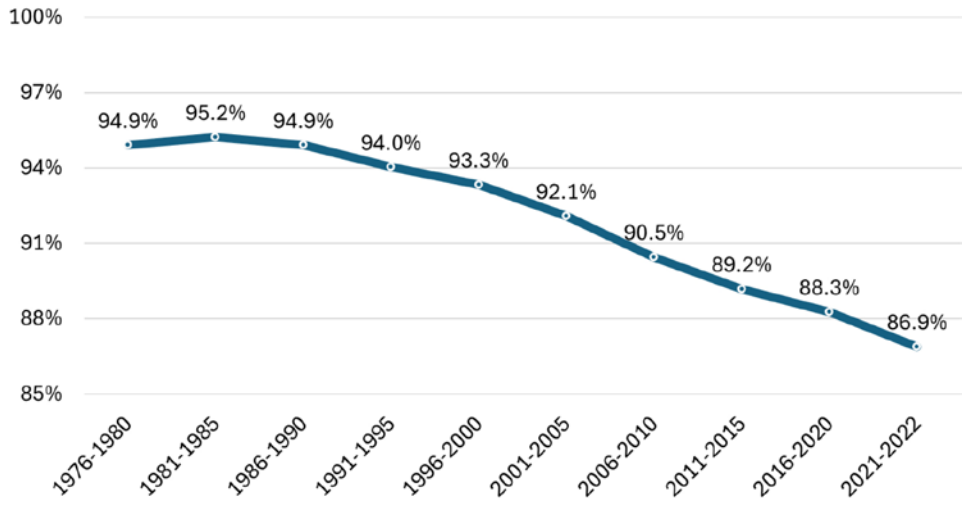
Abbreviations: CRC, colorectal cancer.

*The total number of cancer diagnoses decreased substantially in 2020 due to the COVID-19 pandemic.

**The most recent category includes only two years of diagnoses.

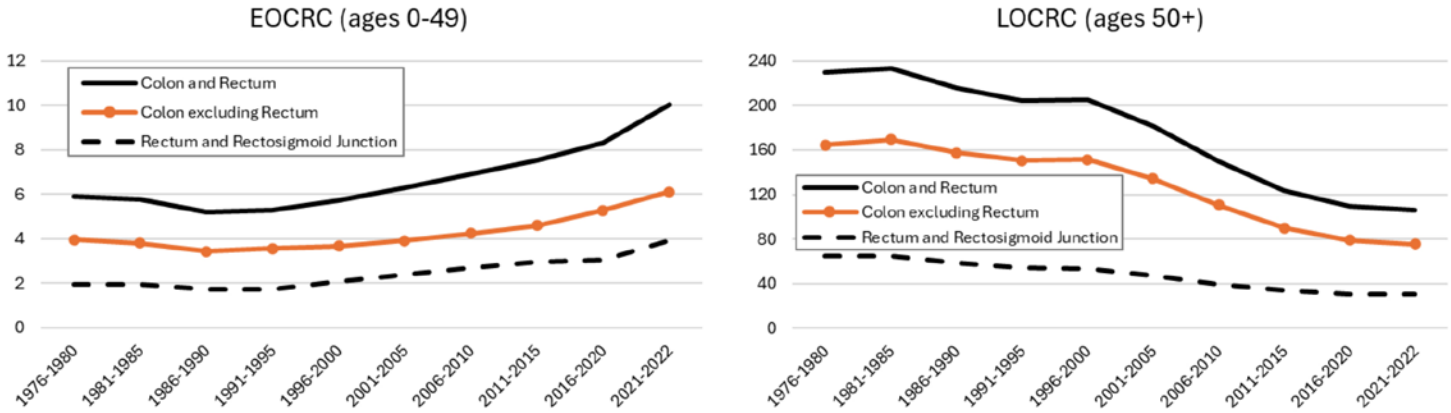
Table 1 displays the average percent changes (APCs) in incidence rates by subtype and age group for 5-year periods from 1976–1980 through 2016–2020 and from 1996–2000 through 2016–2020, excluding the last period that covers only two years (2021–2022). From 1976–1980 through 2016–2020, most subtypes showed increases for younger age groups and decreases for older ages; however, three subtypes (cecum, appendix, and hepatic flexure) showed increases for all age groups and one (splenic flexure) showed increases for almost all ages, with the exception of ages 65–69 and 75–79. APCs were consistently higher among younger individuals (ages 0–49) than older individuals (ages 50+) for all subtypes, with the highest APCs in the youngest individuals (ages 0–29) where calculable. These trends were similar but shifted slightly for the most recent timeframe from 1996–2000 through 2016–2020. Appendix was the only subtype with increasing incidence across all age groups, whereas most other subtypes showed increases among the younger age group (ages 0–49) only, with the highest APCs for appendix, descending colon, rectum, and sigmoid colon cancers in this age group. APCs again tended to be highest in the youngest individuals (ages 0–29) for most subtypes, with the exception of ascending colon and sigmoid colon. These patterns were similar to those for all 22 SEER registries combined from 2001–2002 through 2019–2020, where only appendix cancer showed increasing APCs across all age groups and APCs tended to be higher among younger age groups (Table S3). Although these APCs do not show statistical significance, they display a clear pattern. Overall, APCs tended to be higher among younger versus older age groups for all subtypes of CRC, both in NYS and in all 22 SEER registries combined, indicating that incidence rates of CRC increased faster for younger people in recent decades.

Figure S1. Percent of Late-Onset Colorectal Cancer (LOCRC) to Total Colorectal Cancer (CRC) Cases by Time Period, New York State



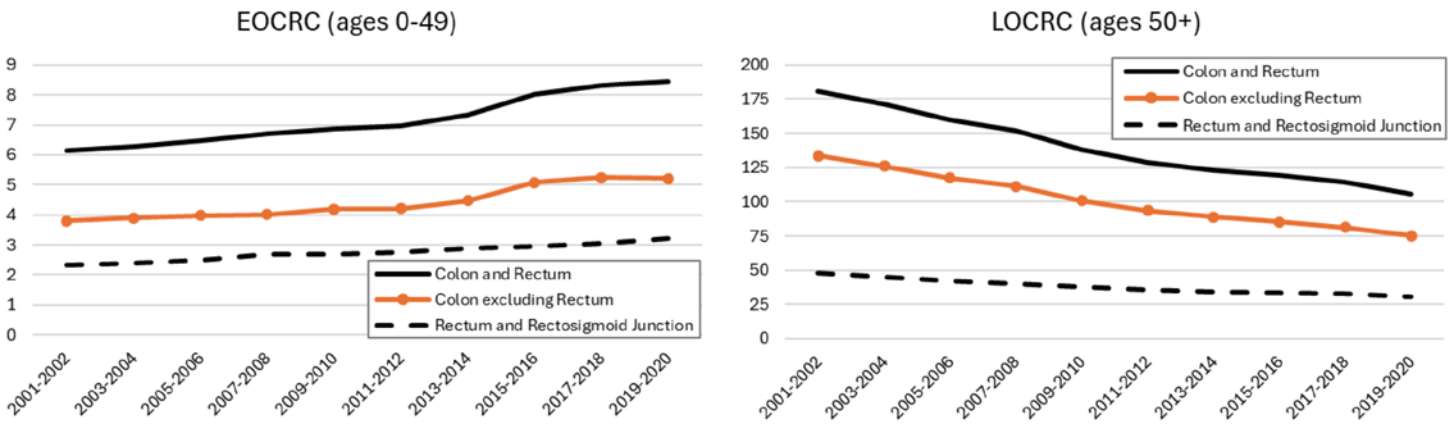
Abbreviations: CRC, colorectal cancer; LOCRC, late-onset colorectal cancer (among ages 50+).

Figure 1. Overall Incidence Rates of Colorectal Cancer (CRC) by Broad Subtype and Age Group, New York State



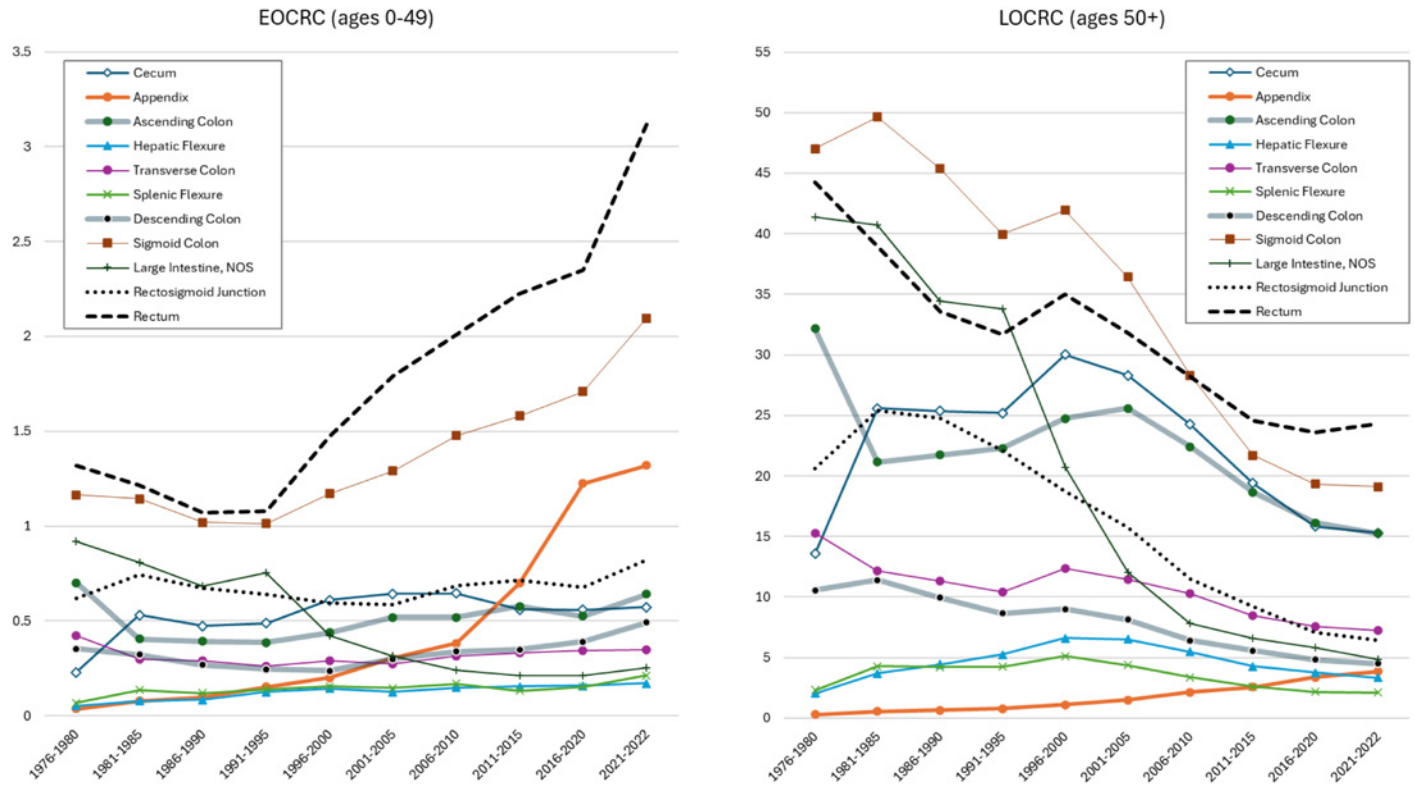
Abbreviations: CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Rates are per 100,000 and age-adjusted to the 2000 US standard population; the y-axis scales differ between the two charts.

Figure S2. Overall Incidence Rates of Colorectal Cancer (CRC) by Broad Subtype and Age Group, 22 Registries



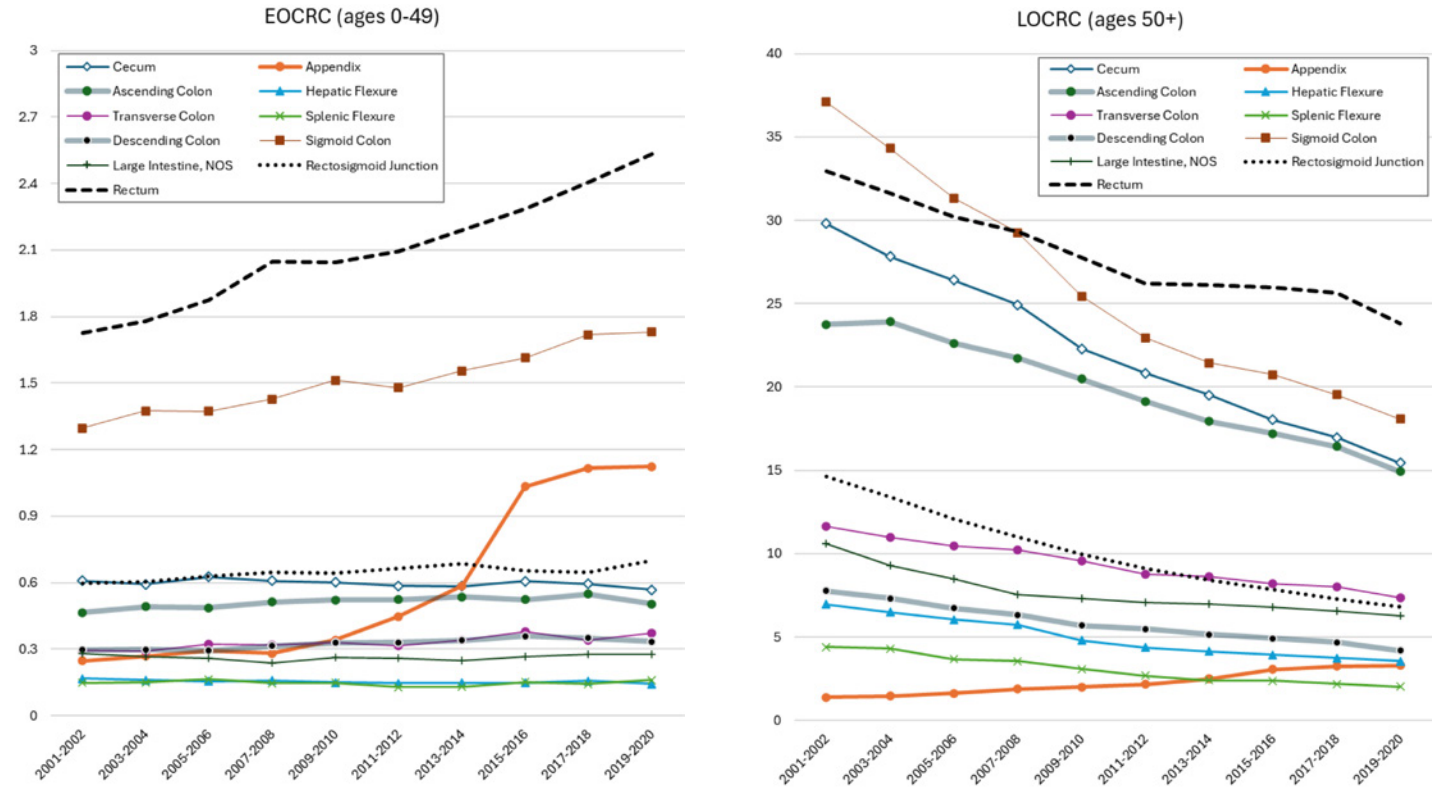
Abbreviations: CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Rates are per 100,000 and age-adjusted to the 2000 US standard population; the y-axis scales differ between the two charts.

Figure 2. Incidence Rates of Colorectal Cancer (CRC) by Subtype and Age Group, New York State



Abbreviations: CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Rates are per 100,000 and age-adjusted to the 2000 US standard population; the y-axis scales differ between the two charts.

Figure S3. Incidence Rates of Colorectal Cancer (CRC) by Subtype and Age Group, 22 Registries



Abbreviations: CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Rates are per 100,000 and age-adjusted to the 2000 US standard population; the y-axis scales differ between the two charts.

Table 1. Average Percent Changes (APCs) by 5-Year Period by Subtype and Age at Diagnosis, New York State

From 1976-1980 through 2016-2020

Age at Diagnosis	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80 - 84	85+	0 - 29	30 - 44	0 - 49	50+	all
Colorectal	11	9	7	4	2	0	-6	-8	-9	-10	-10	-10	-9	19	6	5	-9	-7
Colon	12	8	6	3	1	-1	-6	-8	-9	-10	-10	-9	-9	22	4	4	-8	-7
Rectal	13	14	9	7	4	2	-5	-8	-10	-11	-12	-12	-12	13	8	6	-9	-7
Cecum	22	24	18	13	22	13	7	9	8	4	5	5	5	21	15	18	6	7
Appendix	64	n/a	66	64	43	43	57	38	58	31	40	31	96	79	75	58	38	44
Ascending Colon	-1	13	3	0	-4	-1	-5	-6	-7	-9	-7	-7	-7	4	1	-2	-7	-7
Hepatic Flexure	n/a	35	21	16	25	17	18	16	14	11	10	15	9	n/a	15	16	12	12
Transverse Colon	23	2	2	3	-2	-4	-7	-8	-9	-9	-8	-8	-5	21	-1	-2	-8	-7
Splenic Flexure	n/a	48	6	14	37	19	6	8	-1	4	-2	8	9	56	10	15	4	5
Descending Colon	32	8	17	1	1	2	-8	-11	-11	-9	-10	-10	-9	25	4	2	-9	-8
Sigmoid Colon	16	14	9	6	3	1	-6	-10	-11	-13	-12	-12	-12	12	7	5	-10	-8
Large Intestine, NOS	-4	-11	-18	-15	-13	-16	-19	-21	-21	-23	-21	-21	-17	-9	-16	-15	-20	-20
Rectosigmoid Junction	8	8	0	6	0	-3	-9	-10	-12	-13	-13	-14	-13	11	3	2	-11	-10
Rectum	17	18	13	8	7	5	-3	-6	-8	-10	-11	-11	-12	17	10	8	-7	-5

From 1996-2000 through 2016-2020

Age at Diagnosis	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80 - 84	85+	0 - 29	30 - 44	0 - 49	50+	all
Colorectal	20	15	14	9	5	4	-9	-14	-17	-18	-18	-16	-16	33	11	10	-15	-12
Colon	24	15	13	8	4	0	-10	-15	-17	-18	-17	-16	-16	40	11	10	-15	-13
Rectal	13	16	14	11	7	9	-7	-11	-16	-18	-19	-18	-18	18	12	10	-13	-10
Cecum	31	-6	5	-2	-4	-6	-9	-14	-17	-17	-16	-15	-14	26	-1	-2	-15	-14
Appendix	68	61	58	38	55	31	28	26	43	34	42	45	26	92	49	59	33	42
Ascending Colon	9	6	16	11	-1	6	-6	-8	-11	-12	-11	-10	-9	9	11	5	-10	-9
Hepatic Flexure	97	21	6	14	-4	-4	-2	-12	-15	-13	-14	-14	-12	68	9	3	-13	-12
Transverse Colon	62	14	13	7	5	-5	-7	-15	-14	-12	-10	-11	-11	56	3	5	-11	-10
Splenic Flexure	n/a	33	11	12	-1	-5	-15	-21	-25	-19	-20	-16	-19	75	4	1	-19	-17
Descending Colon	49	38	16	11	13	7	-8	-14	-18	-19	-18	-14	-17	24	14	14	-14	-12
Sigmoid Colon	17	27	14	13	6	4	-11	-17	-20	-23	-21	-21	-22	7	14	10	-17	-15
Large Intestine, NOS	5	-14	-19	-17	-9	-14	-24	-28	-29	-31	-27	-28	-23	-12	-18	-15	-26	-25
Rectosigmoid Junction	15	-1	6	11	1	-1	-18	-19	-25	-26	-25	-24	-25	15	6	4	-22	-19
Rectum	14	26	17	12	10	13	-3	-7	-12	-15	-16	-15	-15	20	15	13	-9	-6

Note: Average percent changes (APCs) are arithmetic averages of the percent changes for each age group by subtype; n/a means that APC is not calculated because of zero incidence in any of the 5-year periods; red and blue indicate positive and negative APC values, respectively, but do not necessarily indicate statistical significance.

Table S3. Average Percent Changes (APCs) by 2-Year Period from 2001-2002 through 2019-2020 by Subtype and Age at Diagnosis, 22 Registries

Age at Diagnosis	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64	65 - 69	70 - 74	75 - 79	80 - 84	85+	00 - 29	30 - 44	00 - 49	50+	all
Colorectal	7	7	4	4	2	2	-3	-5	-6	-7	-7	-7	-7	13	4	4	-6	-5
Colon	11	8	4	4	2	1	-3	-6	-6	-7	-7	-7	-7	17	5	4	-6	-5
Rectal	2	6	4	4	3	3	-2	-3	-6	-7	-7	-7	-8	2	4	3	-4	-3
Cecum	1	0	0	2	-1	-2	-4	-7	-7	-7	-7	-7	-7	1	1	0	-7	-6
Appendix	37	24	22	18	13	13	10	11	14	13	10	10	9	42	20	23	11	15
Ascending Colon	0	16	3	2	1	0	-3	-4	-5	-5	-5	-5	-5	2	3	2	-4	-4
Hepatic Flexure	21	2	5	1	-2	-2	-4	-8	-7	-8	-7	-8	-8	7	1	-1	-7	-7
Transverse Colon	12	0	6	6	1	1	-2	-4	-5	-6	-5	-4	-5	2	4	2	-5	-4
Splenic Flexure	-1	8	6	2	-2	1	-7	-9	-8	-8	-10	-9	-9	-3	3	0	-8	-7
Descending Colon	13	14	4	3	1	0	-2	-5	-6	-8	-8	-6	-8	8	3	2	-6	-5
Sigmoid Colon	9	8	4	5	3	2	-4	-7	-9	-10	-11	-10	-10	7	4	4	-8	-6
Large Intestine, NOS	12	-1	2	1	0	-1	-3	-5	-6	-8	-6	-6	-6	7	0	0	-6	-5
Rectosigmoid Junction	0	2	2	2	1	-1	-6	-6	-10	-10	-11	-11	-11	0	2	1	-8	-7
Rectum	3	7	5	4	4	5	-1	-2	-4	-5	-6	-6	-6	3	5	4	-3	-2

Note: Average Percent Changes (APCs) are arithmetic averages of the percent changes for each age group by subtype; red and blue indicate positive and negative APC values, respectively, but do not necessarily indicate statistical significance.

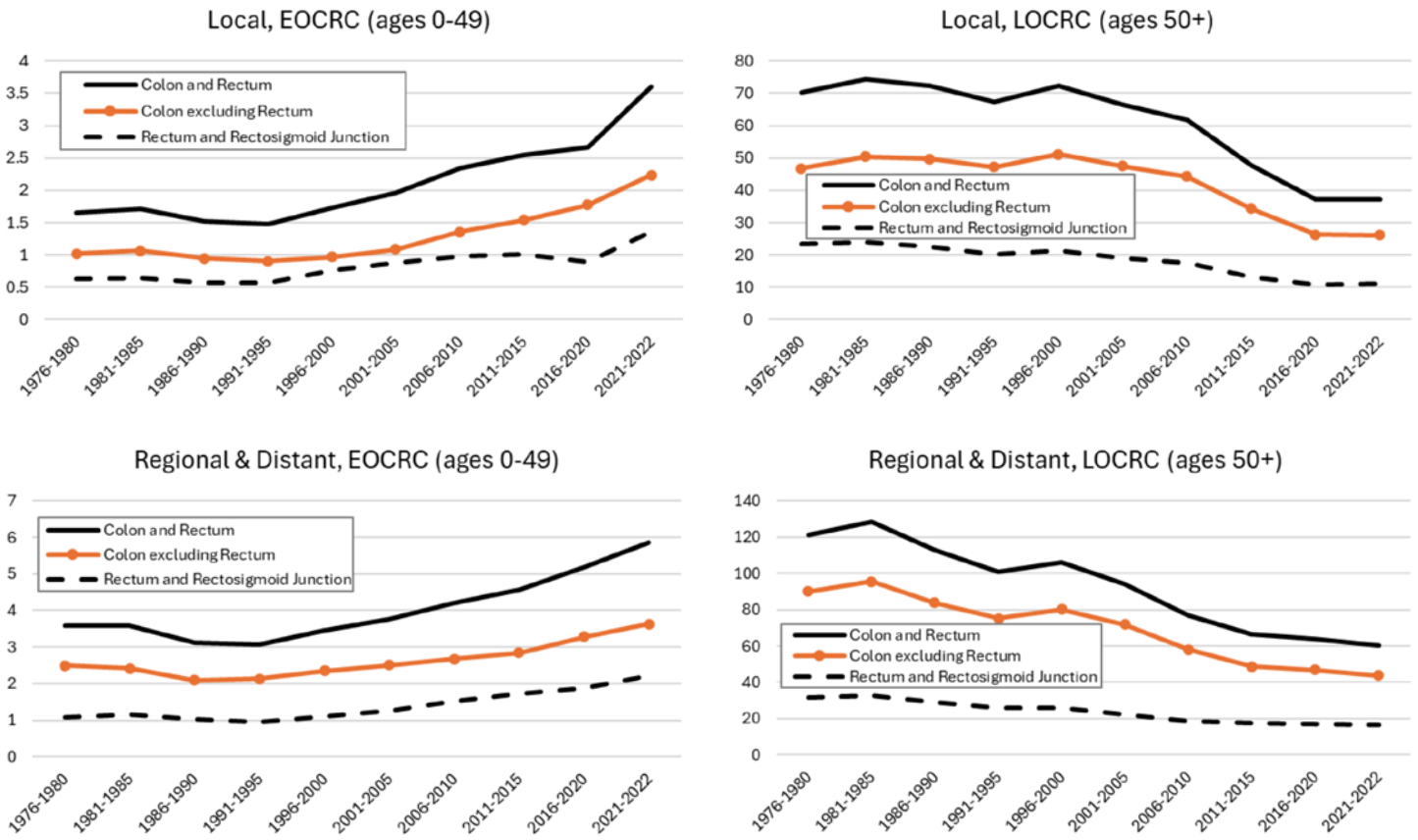
We additionally examined incidence rates of EOCRC and LOCRC by broad subtype and stage category (Figure 3). From 1976–1980 to 2016–2020, incidence of EOCRC increased both for local stage and regional/distant stage cancers, with slightly larger relative increases for local stage colon cancers. For LOCRC, incidence of both local and regional/distant stage cancers decreased, with slightly greater relative decreases for regional and distant cancers. We further examined incidence rates by stage category and detailed subtype (Figure 4). Focusing on the period ending in 2016–2020, due to instability of some estimates for the most recent two-year period, we observed that incidence rates of local stage EOCRC declined for some subtypes from 2006–2010 to 2016–2020, whereas it markedly increased for appendix. In contrast, incidence of regional- and distant-stage EOCRC showed much steeper increases for rectum and sigmoid colon than for appendix cancers beginning in 1991–1995. For LOCRC, incidence of all subtypes except appendix cancer decreased during the two decades from 2001–2005 to 2016–2020 for both local and regional/distant stage cancers. Subtypes with higher incidence rates decreased more steeply than those with lower incidence rates and, among the subtypes with higher incidence rates, rectal cancers decreased relatively less and thus had the highest incidence in the past decade for both local and regional/distant cancers.

We also analyzed the percentage of local-stage diagnoses to total diagnoses (local, regional, and distant stages combined) for both EOCRC and LOCRC (Figure S4), to assess the stage distribution by subtype and age group over time. We observed a higher percentage of local-stage diagnoses for LOCRC versus EOCRC for almost all time points and subtypes, with the exception of appendix where the opposite was true: The percentage of local-stage appendix cancer was consistently higher for EOCRC than LOCRC. Further, for all CRC combined, we observed that the percent of local-stage diagnoses decreased since the 2006–2010 period for LOCRC, whereas this percentage increased for EOCRC colon cancer, in part due to increases in the percentage of appendix cancers diagnosed at an early stage.

Discussion

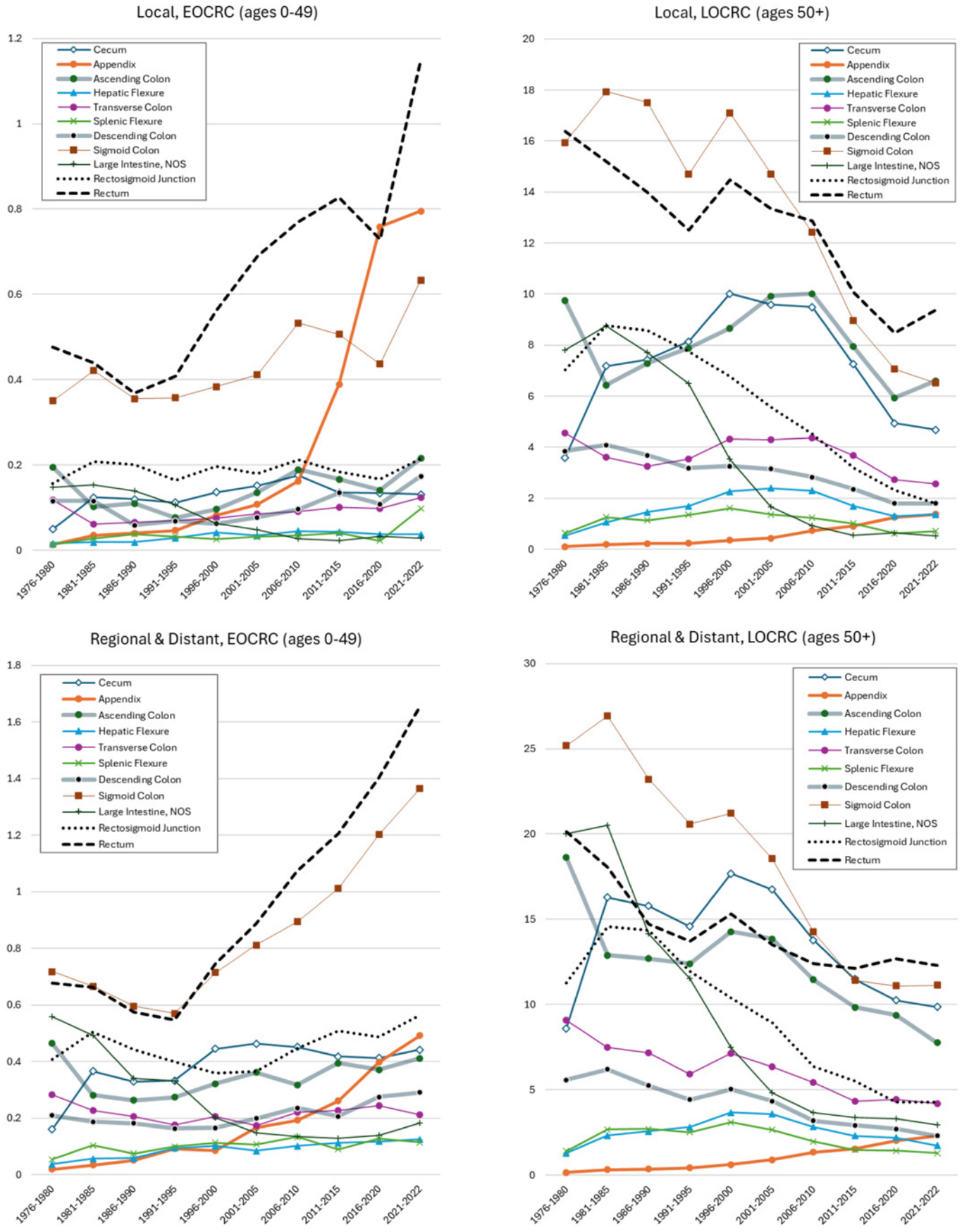
In our study of NYS CRC incidence over an extensive period, we observed results that are consistent with previous evidence of increasing CRC incidence in individuals under age 50. Our analysis expands on prior studies and indicates that increasing incidence of EOCRC is attributable to increases in rectal, sigmoid, and appendix cancers. In addition, the results show increasing trends in appendix cancer among older individuals despite the decreasing incidence of LOCRC overall. APCs further indicate that CRC incidence rates have increased faster (or decreased more slowly) for

Figure 3. Incidence Rates of Colorectal Cancer (CRC) by Broad Subtype, Age Group, and Stage Category, New York State



Abbreviations: CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Rates are per 100,000 and age-adjusted to the 2000 US standard population; the y-axis scales differ between the four charts.

Figure 4. Incidence Rates of Colorectal Cancer (CRC) by Subtype, Age Group, and Stage Category, New York State



Abbreviations: CRC, colorectal cancer; EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Rates are per 100,000 and age-adjusted to the 2000 US standard population; the y-axis scales differ between the four charts.

younger versus older individuals across CRC subtypes, and that some subtypes (appendix, descending colon, rectum, and sigmoid colon) have had greater increases than the other subtypes among younger individuals. We observed similar findings in the SEER 22 data (2001–2020). These findings highlight the importance of conducting analyses by CRC subtype, as well as by age group.

Though many studies have reported increasing incidence of EOCRC and some possible contributing factors, only a few studies have examined these trends by some grouping of CRC subtypes. For example, CRC incidence rates were examined for colon and rectum by Chen and coauthors,[5] for colon with appendix versus colon without appendix by Montminy and coauthors,[7] for right colon, left colon, and rectum by Patel and coauthors,[4] and for five subtypes (anus, appendix, distal colon, proximal colon, and rectum) and histology by Meyers and coauthors.[6] We examined CRC incidence rates for 11 subtypes and produced more detailed information about the CRC trends and, specifically, which subtypes are driving the increasing incidence rate of EOCRC. While previous studies[3,8] noted the diverging trends of appendix cancer among the CRC subtypes, we further found that rectal and sigmoid colon cancers, as well as appendix cancer, contributed to increasing EOCRC incidence.

A key finding of this analysis is that the increase in EOCRC incidence has been driven primarily by three subtypes (rectum, sigmoid colon, and appendix). Though the other subtypes (except for "large intestine, NOS") have also been increasing (albeit gradually) among the younger age group, their contribution to the increase in EOCRC incidence is low. These results suggest that research on the causes of increasing EOCRC incidence should focus on the subtype-specific trends and in particular on possible risk factors for tumors in the rectum, sigmoid colon, and appendix.

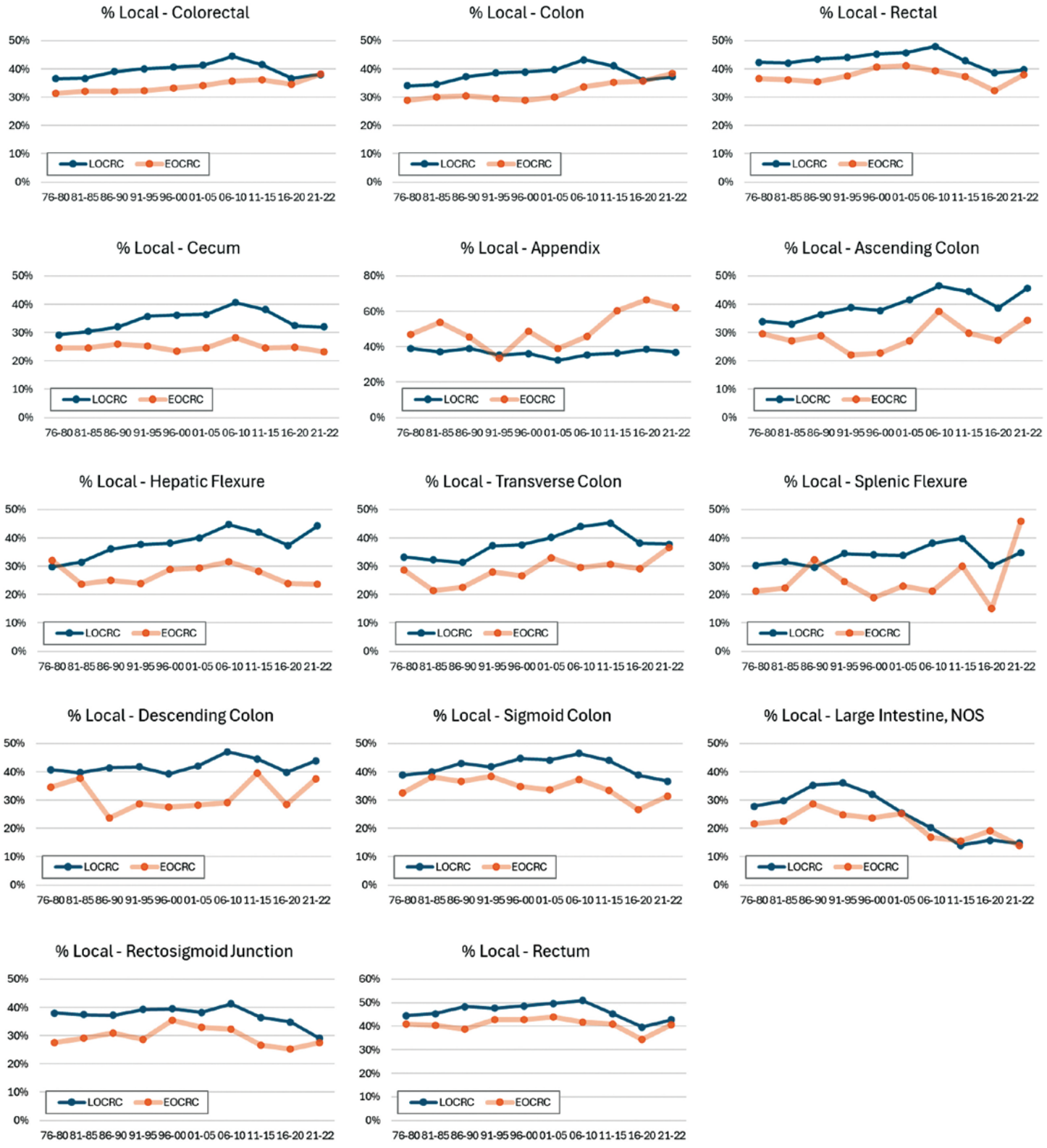
Another key finding is that appendix cancer has increased in incidence for both younger and older individuals, even though incidence of all other subtypes has decreased or has minimally changed among the older age group. Since 1996–2000, appendix was the only subtype of LOCRC with increasing APCs. APCs for appendix cancer were higher than any other CRC subtypes across all age groups, with higher values for younger age groups. Incidence rates of appendix cancer appear to have increased faster in more recent years among individuals under age 50.

These results suggest that additional work is needed to find possible causes for increasing incidence of appendix cancers in both younger and older individuals and to improve understanding of the factors contributing to specific subtypes of colorectal cancer among different age groups. Although studies have identified risk factors for EOCRC, such as diet, smoking, alcohol consumption, lack of physical activity, obesity and overweight, genetic predisposition, family history, type 2 diabetes, and prior radiation exposure,[3-6,11] these factors also apply to CRC in general and do not help explain the rapid increase in incidence of certain subtypes of EOCRC in recent decades.

Screening is a key factor in two respects. First, it explains the differing trends between EOCRC and LOCRC, in that screening has contributed to the decrease in CRC incidence in the older population, through the detection and removal of adenomatous polyps,[11,12] while individuals younger than 45 are not eligible for routine screening and testing is available only for high-risk or symptomatic individuals. The United States Preventive Services Task Force (USPSTF) previously recommended routine CRC screening for all adults aged 50–75, but in 2021 expanded the recommendations to include adults aged 45–49.[13] Among age eligible adults in NYS, those who met the USPSTF recommendations for CRC screening increased from about 40% in 1996 to approximately 70% in recent years, with higher percentages among adults aged 65–75.[14] Second, screening partly explains why incidence of appendix cancers has increased despite decreasing incidence of LOCRC overall, in that appendix cancers are not consistently detectable with screening but are mostly diagnosed incidentally during other medical procedures.[15] However, screening cannot explain the results observed in younger individuals. In analyses stratified by stage at diagnosis, we observed generally similar trends for both local and regional/distant stage cancers in both the younger and older populations. This suggests that the increases in incidence of EOCRC are not due to incidental or screen-detected early-stage cancers and reflect true increases in disease in younger individuals.

One limitation of this analysis is that some changes in the incidence trends are likely attributable in part to changes in the coding guidelines and do not represent true changes in cancer incidence. For example, around 1995, the incidence of "large intestine, NOS" dropped steeply while the incidence of other subtypes slightly increased. This corresponded to a coding change where adenocarcinoma in a polyp present in more than one segment of the colon was reported as multiple primaries by United States cancer registries, whereas prior to 1995 it was reported as one primary tumor and coded as "Large intestine, NOS".[16,17] Additionally, starting in 2015, an exception was added to the cancer coding manual so that carcinoid tumors of the appendix were regarded as malignant and thus reportable.[8,18] These changes in the coding guidelines may have contributed to the increasing incidence rates of appendix cancer in the 2010s. Yet, it is also important to note that the incidence rate of appendix cancer had already been increasing steadily, which implies a true increase in appendiceal cancer incidence rates.[8] While Bleyer and coauthors suspected that much of the reported increase in EOCRC was caused by the reclassification of appendiceal carcinoids and neuroendocrine tumors as malignant,[19] our results indicate that EOCRC has been increasing in incidence for all subtypes, and that rectal and sigmoid colon cancers are responsible for a large proportion of late-stage EOCRC. Strengths of this analysis include the large number of subtypes examined and the large amount of data included, both for NYS, which has the 4th highest number of CRC cases per year of all US states,[2] and for the 22 SEER registries combined.

Figure S4. Percentage of Local-Stage Diagnoses by Subtype, Age Group, and Stage Category, New York State



Abbreviations: EOCRC, early-onset colorectal cancer; LOCRC, late-onset colorectal cancer.
Note: Percentage of local means percent of diagnoses at the local stage relative to the total diagnoses at all three stages (local, regional, and distant); the y-axis scales differ across the charts.

In summary, this paper expands upon previous work showing trends of increasing EOCRC and decreasing LOCRC and provides more details of CRC incidence by age group, subtype, and stage, highlighting that studies of trends in CRC incidence should consider results by subtype. To help improve understanding of the recent trends in CRC incidence, additional research is needed to clarify: (1) why the incidence rates of rectum, sigmoid colon, and appendix cancers have increased rapidly in recent decades while the other subtypes have remained relatively low among the younger population; and (2) why appendix cancer incidence has increased across all age groups, with the fastest increase among the younger population. With the shift toward EOCRC comprising an increasing percentage of CRC and the evidence that screening has contributed to decreasing LOCRC,[11,12] additional work is needed to better understand and reverse the recent trends of increasing CRC incidence in the younger population.

References

1. New York State Cancer Registry. Cancer Incidence and Mortality in New York State, 1976–2022. Accessed Apr 2025. **http://health.ny.gov/statistics/cancer/registry/**
2. American Cancer Society. Cancer Facts & Figures 2025. Atlanta: American Cancer Society; 2025. Accessed Apr 2025. **http://cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2025/2025-cancer-facts-and-figures-acs.pdf**
3. Sung H, Siegel RL, Laversanne M, Jiang C, Morgan E, Zahwe M, Cao Y, Bray F, Jemal A. Colorectal cancer incidence trends in younger versus older adults: An analysis of population-based cancer registry data. *Lancet Oncol* Jan 2025; 26(1):51–63. doi: 10.1016/S1470-2045(24)00600-4.
4. Patel SG, Karlitz JJ, Yen T, Lieu CH, Boland CR. The rising tide of early-onset colorectal cancer: A comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. *Lancet Gastroenterol Hepatol* Mar 2022;7(3):262–274. doi: 10.1016/S2468-1253(21)00426-X.
5. Chen HLR, Chong QD, Tay B, Zhou S, Wong EYT, Seow-En I, Tan KK, Wang Y, Seow A, Tan KE, Tan BHI, Tan SH. Trends in early-onset colorectal cancer in Singapore: Epidemiological study of a multiethnic population. *JMIR Public Health Surveill* Feb 2025; 11:e62835. doi: 10.2196/62835.
6. Meyers AL, Dowty JG, Mahmood K, Macrae FA, Rosty C, Buchanan DD, Jenkins MA. Age-specific trends in colorectal, appendiceal, and anal tumour incidence by histological subtype in Australia from 1990 to 2020: A population-based time-series analysis. *medRxiv* [Preprint]. Apr 2025: 2025.04.21.25326138. doi: 10.1101/2025.04.21.25326138.
7. Montminy EM, Zhou M, Long C, Wani S, Patel SG, Karlitz JJ. Impact of carcinoid and appendiceal site inclusion on early onset colon cancer incidence rate analysis in the United States. *Clin Gastroenterol Hepatol* Oct 2023; 21(11):2972–2974.e1. doi: 10.1016/j.cgh.2022.10.010.
8. National Cancer Institute – Surveillance, Epidemiology, and End Results (SEER) Program. 22 cancer registries include 13 state cancer registries (Connecticut, Hawaii, Idaho, Iowa, Illinois, Kentucky, Louisiana, Massachusetts, New Mexico, New Jersey, New York, Texas, and Utah) and 9 targeted cancer registries (Alaska Native Tumor Registry, Atlanta, Greater Georgia, Rural Georgia, San Francisco-Oakland, San Jose-Monterey, Greater California, and Los Angeles). Accessed Sep 2025. **http://seer.cancer.gov/registries/terms.html**
9. National Cancer Institute – Surveillance, Epidemiology, and End Results (SEER) Program. Site Recode ICD-O-3/WHO 2008 Definition. Accessed Dec 2025. **http://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html**
10. National Cancer Institute – Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Software. Latest Release: Version 9.0.41 - July 29, 2025. Accessed Sep 2025. **http://seer.cancer.gov/seerstat/**
11. Lin JS, Perdue LA, Henrikson NB, Bean SI, Blasi PR. *Screening for Colorectal Cancer: An Evidence Update for the U.S. Preventive Services Task Force*. Evidence Synthesis No. 202. AHRQ Publication No. 20-05271-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; 2021.
12. Jacobsson M, Wagner V, Kanneganti S. Screening for Colorectal Cancer. *Surg Clin North Am* 2024 Jun;104(3):595–607. doi: 10.1016/j.suc.2023.11.009.
13. US Preventive Services Task Force; Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Krist AH, Kubik M, Li L, Ogedegbe G, Owens DK, Pbert L, Silverstein M, Stevermer J, Tseng CW, Wong JB. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. *JAMA* May 2021; 325(19):1965–1977. doi: 10.1001/jama.2021.6238. Erratum in: JAMA. Aug 2021; 326(8):773. doi: 10.1001/jama.2021.12404.
14. New York State Department of Health, Division of Chronic Disease Prevention, Bureau of Chronic Disease Evaluation and Research. Colorectal Cancer Screening, New York State Adults Ages 45–75, 2023, 2022, 2020, 2019, 2017, 2016, 2015, 2014, 2013, 2012, 2010, 2006, and 1996–2002. BRFSS Brief. Accessed Dec 2025. Source: **http://health.ny.gov/statistics/brfss/reports/**
15. Van de Moortele M, De Hertogh G, Sagaert X, Van Cutsem E. Appendiceal cancer: A review of the literature. *Acta Gastroenterol Belg* 2020 Jul-Sep;83(3):441–448.
16. Standards of the Commission on Cancer. *Volume II: Registry Operations and Data Standards (ROADS)*. Accessed Sep 2025. **http://facs.org/media/wxkot24a/roads1996.pdf**
17. Dickie, L., Johnson, CH., Adams, S., Negoita, S. (November 2024). *Solid Tumor Rules*. National Cancer Institute, Rockville, MD 20850. Accessed Sep 2025. **http://seer.cancer.gov/tools/solidtumor/current/STM_Combined.pdf**
18. Adamo M, Dickie, L, Ruhl J. (January 2015). *SEER Program Coding and Staging Manual 2015*. National Cancer Institute, Bethesda, MD 20850-9765. Accessed Sep 2025. **http://seer.cancer.gov/archive/manuals/2015/SPCSM_2015_maindoc.pdf**
19. Bleyer A, Ries LAG, Cameron DB, Mansfield SA, Siegel SE, Barr RD. Colon, colorectal, and all cancer incidence increase in the young due to appendix reclassification. *J Natl Cancer Inst* 2025 Jul 1;117(7):1340–1349. doi: 10.1093/jnci/djaf038.

Original Article

Leveraging Large Language Models for Real-World Data Evidence: A Framework for Automated Treatment Extraction and Data Harmonization

Abhishek Shivanna, PhD[a]; Austin Fitts, PharmD[b]; Jordan Tschida, PhD[a]; Mayanka Chandrashekar, PhD[a]; Amie Hwang, PhD[b]; Elizabeth Hsu, PhD[b]; John Gounley, PhD[a]; Heidi A. Hanson, PhD[a]

Abstract: Background: The ability to comprehensively collect treatment information from cancer patient medical records would enable studies to evaluate real-world benefits and risks tied to specific treatments. Currently, it is difficult to systematically collect high-quality treatment information because it is often stored in unstructured text. Manually extracting and standardizing drug and regimen data is time-intensive. Recent advances in large language models (LLMs) offer a potential solution for automated extraction of structured treatment information from clinical text. Objective: This study systematically evaluates the utility of four LLMs from the Llama family for automated extraction of oncology treatment information from clinical text. This information can guide researchers using cancer registry data to provide insights into cancer care and outcomes beyond clinical trials. Methods: Four instruction-tuned Llama models with varying parameter counts (1B, 3B, 8B, and 70B) were evaluated for their ability to extract treatment information from clinical documents. A unified oncology knowledge base integrating seven major public data sources was developed to standardize and normalize extracted entities—a critical step for harmonizing data from diverse sources. Extracted treatment data were compared against expert-annotated ground truth. Model performance was assessed using accuracy metrics (Precision, Recall, F1-Score) and operational feasibility metrics, including processing speed and structural compliance of the output. Results: A strong positive correlation was observed between model size and extraction accuracy. F1-score improved from 0.609 for the 1B model to 0.710 (3B), 0.807 (8B), and 0.828 (70B). While larger models demonstrated superior accuracy and compliance, they incurred higher computational costs. The modest performance difference between 8B and 70B suggests diminishing returns with increasing model size. Conclusions: LLMs represent a viable technology for automating oncology treatment extraction. The 8B-parameter model emerged as a highly effective option, balancing high accuracy and computational efficiency. Selecting an appropriate LLM for deployment in cancer registries involves a trade-off between desired accuracy and available operational resources. Harmonizing extracted entities with the oncology knowledge base facilitates standardized integration into common data models, enhancing data quality for real-world evidence analyses.

Key words: *Artificial intelligence, natural language processing, large language models, cancer registry, data normalization, data extraction, oncology informatics*

Introduction

Accurate and timely capture of cancer treatment data is central to the mission of cancer registries. Beyond their traditional roles, these data serve as a cornerstone for real-world evidence in oncology studies. By providing robust insights into treatment patterns, cancer registries data can be used to advance clinical research, shape public health policy, and drive evidence-based decision making in oncology care. Currently, there is no systematic way to extract treatment information from heterogeneous unstructured and semi-structured documents (e.g., pathology reports, clinician notes, treatment summaries). Persistent gaps and discordance in registry treatment fields, incomplete capture of systemic therapy and radiation, and under-ascertainment of radiotherapy offer further reasons to consider automation to augment abstraction workflows.[1–4] The utilization of large language models (LLMs) for automatic extraction of treatment information is a promising research direction. However, extraction alone does not address the data harmonization issue, which arises from inconsistencies and variations in how treatments are documented across different sources. Harmonized data is important

[a]Advanced Computing for Health Sciences, Oak Ridge National Laboratory, Oak Ridge, TN. [b]National Cancer Institute, National Institutes of Health, Washington, DC.

Address correspondence to Abhishek Shivanna, Advanced Computing for Health Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; shivannaa@ornl.gov

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer program established by the US Department of Energy (DOE) and the National Cancer Institute of the National Institutes of Health. This work was performed under the auspices of the DOE by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

for real-world evidence (RWE) in oncology because it ensures that data from diverse sources can be meaningfully compared, integrated, and analyzed. The purpose of this paper is to provide a blueprint for cancer registry researchers interested in extracting treatment information from their own records. We compare the utility of different sized LLMs for extracting treatment information and then provide a solution for harmonizing the extracted information.

Recent advances in artificial intelligence, particularly open-weight large language models (LLMs), offer a promising path toward automating complex information extraction. The Llama 3 family and its predecessors introduced strong instruction following and long-context capabilities, building on reinforcement learning from human feedback and an open model ecosystem that supports reproducible evaluation.[5–7] In clinical information extraction (IE), LLMs have shown competitive performance in zero/few-shot settings and interpretable feature construction from notes.[8,9] This trajectory complements a decade of community shared tasks and benchmark corpora that established evaluation practices and common error profiles for medication and relation extraction (e.g., MADE 1.0, 2018 n2c2 ADE/medication extraction, DrugProt), alongside surveys synthesizing advances in named entity recognition that underpin these pipelines.[10–13] Methodologically, techniques such as retrieval-augmented generation, self-consistency, and principled prompt design can improve reliability for high-stakes extraction.[14–17]

To address normalization challenges inherent in automated extraction, we developed a comprehensive oncology knowledge base that integrates multiple authoritative sources. Anchoring extraction to controlled vocabularies and classic clinical NLP infrastructure (UMLS, cTAKES, MetaMap) enables robust concept mapping for downstream use.[18–20] Complementary rule-based and hybrid IE systems such as MedXN have demonstrated high-fidelity medication extraction with RxNorm normalization, and two-stage pipelines have shown practical harmonization of drug mentions into observational databases.[21,22] At the analytic level, RxNorm–ATC/DDD crosswalks facilitate scalable assessment of prescriptions and regimens.[23]

This study provides a systematic comparison of four instruction-tuned Llama models spanning 1B, 3B, 8B, and 70B parameters for extracting oncology treatment information from clinical text. Our primary objective is to quantify trade-offs between model scale, extraction accuracy, and operational efficiency. We evaluate models against a standardized, expert-annotated task suite and situate outputs within the unified, multisource oncology knowledge base described above to support normalization and downstream integration for generating high-quality real-world datasets for population-scale analyses of cancer treatment outcomes.

Methods

Study Design and Overview

We conducted a comparative evaluation of instruction-tuned large language models (LLMs) for the extraction of oncology drug mentions from clinical text. The objective was to quantify extraction accuracy across model sizes using a two-stage pipeline. The first stage involved model inference using structured prompts designed to produce JSON-formatted outputs exclusively. The second stage applied drug-name normalization to align extracted mentions with a canonical vocabulary. Models were not fine-tuned on task-specific data but operated with fixed parameters guided solely by in-context instructions.

Data Sources and Preparation

Dataset: We evaluated the pipeline using a curated set of 863 oncology abstract records from the University of Southern California hospital systems, each annotated by domain experts for systemic treatment information. Each record corresponds to a single patient summary containing demographic, diagnostic attributes, and free-text treatment descriptions. Key structured fields included age at diagnosis and primary cancer site, while three narrative fields summarized systemic treatments: chemotherapy, hormone therapy, and immunotherapy notes. These narratives were concatenated into a single string for model input.

Oncology Treatment Normalization Database: We compiled a custom multisource resource to normalize drug and regimen names across brand/generic variants, abbreviations, and misspellings. This database integrates seven sources: HemOnc.org (HemOnc),[24,25] NCI Cancer Medications (CanMED),[26] DrugBank Online (DrugBank),[27] U.S. National Library of Medicine RxNorm (RxNorm),[28,29] National Cancer Institute Thesaurus (NCIt),[30] SEERRx: Interactive Antineoplastic Drugs Database (SEERRx),[31] and Aggregate Analysis of ClinicalTrials.gov (AACT).[31,32] We harmonized entries by collapsing synonyms to preferred generic names and created composite entries for multi-drug regimens linked to constituent agents. The *DrugNormalizer* module performs case-insensitive dictionary matching, synonym/brand resolution, acronym expansion, and fuzzy matching to canonical generic strings. This oncology treatment normalization database is publicly available for use by the research community.[33]

Models and Implementation

Models: We evaluated four instruction-tuned LLMs from the Llama 3 family (Meta AI):

Llama-3.2-1B-Instruct, *Llama-3.2-3B-Instruct*, *Llama-3.1-8B-Instruct*, and *Llama-3.3-70B-Instruct*. These span 1B, 3B, 8B, and 70B parameters, respectively, enabling size-wise comparisons under identical prompting and post-processing.

Prompt and output constraints: Models received a prompt instructing the extraction of drug mentions explicitly normalized to generic names based on prior knowledge available within the model. The prompt explicitly instructed the models to return only JSON-formatted outputs, providing two in-context examples, and specifically directed the models to refrain from including any explanatory or additional text. The exact prompt used was:

```
   Extract drug names from the clinical
text following these rules:
1. Identify all explicit drug mentions
   in the text.
2. Handle common cases:
   - Abbreviations: Convert to full
     generic names (e.g., Ara-C →
     cytarabine).
   - Misspellings: Correct common errors
     (e.g., Methotrxate → methotrexate).
   - Brand names: Convert to generics
     (e.g., Oncovin → vincristine).
3. Unknown generics: If a generic
   equivalent is unknown, include the
   raw drug name.
4. Return ONLY the JSON object, no
   explanations or code examples.
      Examples:
      Text: 'Patient received MTX,
Cisplatin, and Ara-C'
{
       "drugs": ["methotrexate",
  "cisplatin", "cytarabine"]
}
      Text: 'Doxorubicin, L'asp, XL184
and daunorubicin were administered'
{
       "drugs": ["doxorubicin",
  "asparaginase", "cabozantinib",
  "daunorubicin"]
}
      Process this text:
      [text_concat]
      Return ONLY valid JSON:
```

Compute and inference procedure: Inference used a compute node with 8× NVIDIA A100 GPUs. Cases were batch-processed, enforcing strict JSON compliance through a lightweight post-processor.

Normalization and comparison to gold standard: Predicted and gold-standard drug annotations were normalized to their canonical generic drug names using the *Drug-Normalizer*, and comparisons were conducted at the level of these standardized generic drug names.

Evaluation Metrics

Let T denote the set of true (gold) treatments and E the set of extracted treatments after normalization for a given case. We computed:

- Precision = $\frac{(|E \cap T|)}{(|E|)}$
- Recall = $\frac{(|E \cap T|)}{(|T|)}$
- F1-Score = $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

Metrics were computed per case and macro-averaged across cases to weight each patient record equally. We report mean ± standard deviation for each model. Ninety-five percent confidence intervals (CIs) for the mean were estimated as mean ± 1.96 × (SD / √N), where N is the number of cases.

Additionally, we computed two metrics to quantify extraction errors:

Missing drugs (false negatives): Gold-standard drug entities not captured by the model after normalization.

Hallucinated drugs (false positives): Predicted drug entities (after normalization) that are not present in the expert-annotated gold standard for that record. Because the gold standard emphasizes explicit drug mentions, this category can include (i) drugs truly absent from the source text and (ii) clinically plausible regimen/protocol expansions or contextually inferred agents that were not explicitly documented/annotated.

Results

This study systematically assessed the performance of four instruction-tuned Llama models (1B, 3B, 8B, and 70B parameters) in extracting standardized oncology treatment information from clinical text. Models were evaluated based on accuracy metrics (precision, recall, and F1-score), the impact of entity normalization, instruction compliance, processing speed, error profiles, and performance variations across text lengths.

Overall extraction performance (normalized)

Macro-averaged precision, recall, and F1 increased monotonically with model size (Figure 1). The 70B model achieved the highest F1 (0.828), with 8B close behind (0.807; Δ=0.021). Case-level 95% CIs were narrowest for 8B/70B (Figure 2), indicating more consistent performance across notes. Additionally, precision, recall, and F1-score distributions for the 8B and 70B models were skewed towards higher values, further highlighting strong performance consistency (Figure 3).

Effect of normalization

Normalization improved enhanced extraction performance across all models (Table 1). F1 gains were largest for 1B (+0.120; 24.5% relative to pre-normalization) and tapered with scale (3B: +0.088; 8B: +0.056; 70B: +0.032), consistent with systematic correction of abbreviations, brand/generic variants, and spelling noise.

Instruction compliance

Instruction compliance varied sharply (Figure 4). The 70B model was perfectly compliant (100%), whereas 8B was the least compliant (13.7% compliant), consistent with the performance–latency scatter (Figure 5). Interestingly, despite its larger size, the 70B model processed cases faster

Figure 1. Macro-averaged Precision, Recall, and F1 after Normalization across Models (1B–70B)

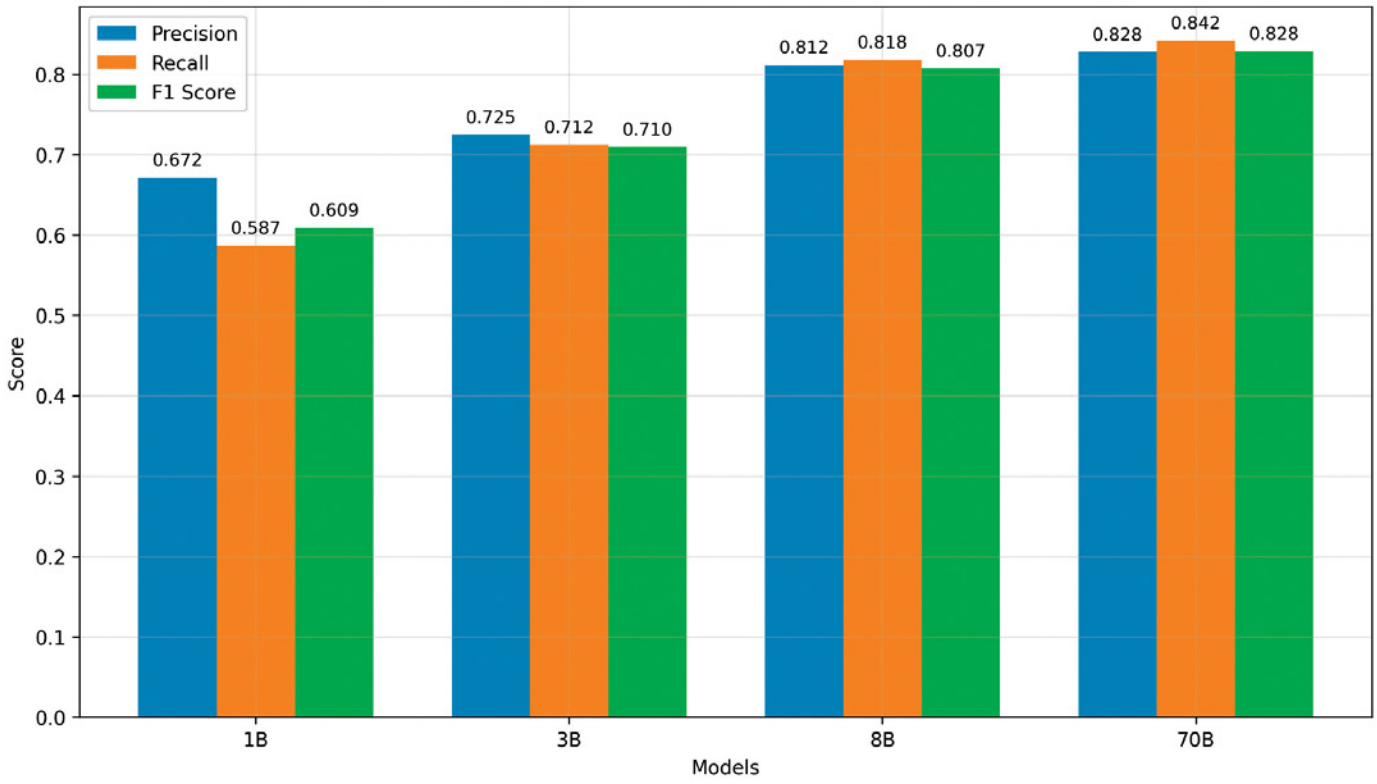


Figure 2. Mean Precision, Recall, and F1 with 95% CIs (Computed across Cases)

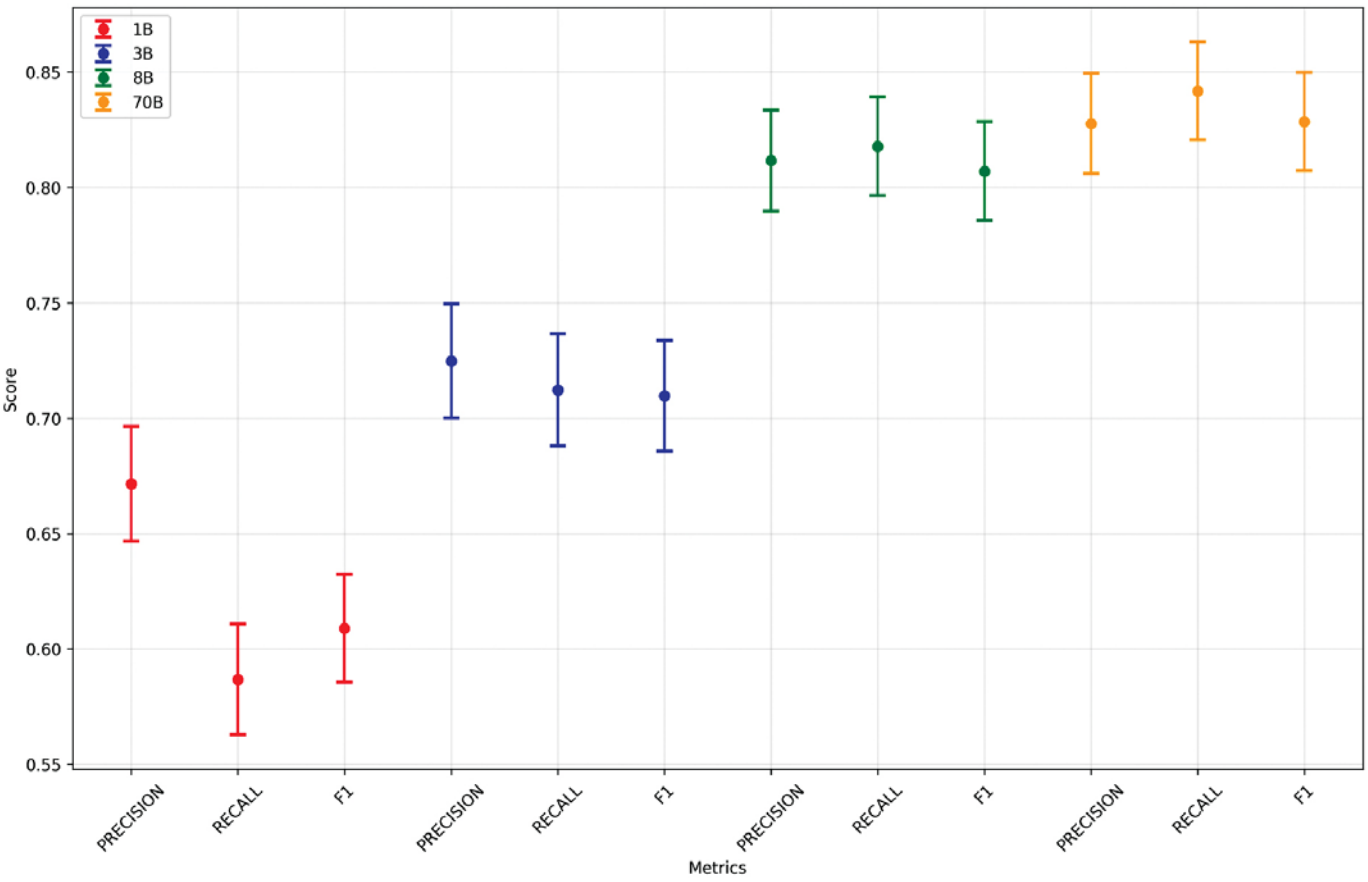


Figure 3. Violin Plots of Case-Level Precision, Recall, and F1 after Normalization

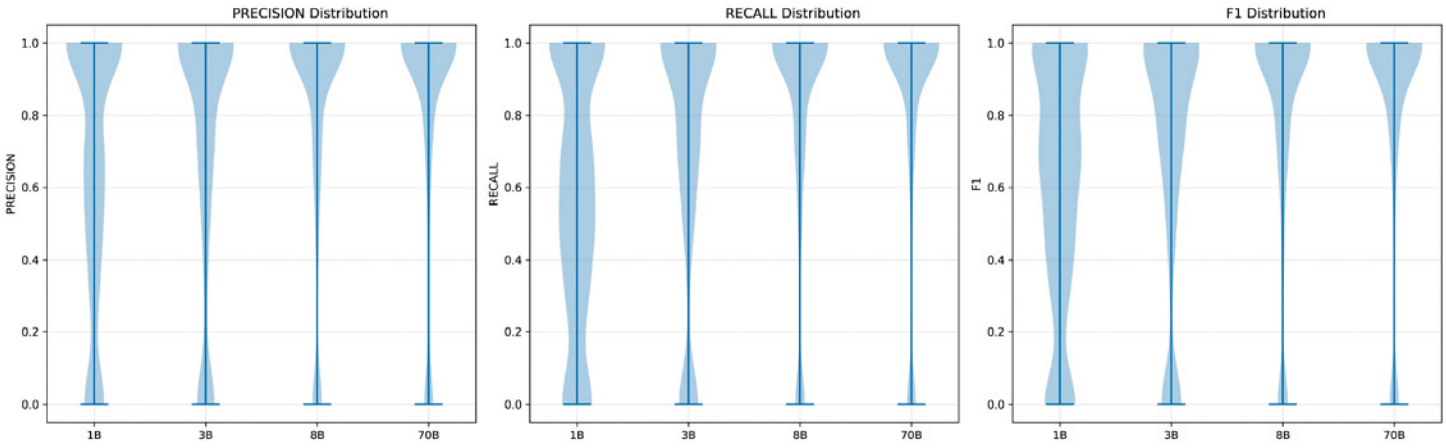


Table 1. Before vs. After Normalization (Macro-Averaged Across Cases). Δ Denotes (Post Pre)

	Precision			*Recall*			*F1*		
Model	**Pre**	**Post**	**Δ**	**Pre**	**Post**	**Δ**	**Pre**	**Post**	**Δ**
1B	0.547	0.672	+0.125	0.466	0.587	+0.121	0.489	0.609	+0.120
3B	0.635	0.725	+0.090	0.624	0.712	+0.088	0.622	0.710	+0.088
8B	0.754	0.812	+0.058	0.761	0.818	+0.057	0.751	0.807	+0.056
70B	0.795	0.828	+0.033	0.809	0.842	+0.033	0.796	0.828	+0.032

Figure 4. Instruction Compliance Rates for JSON-only Outputs

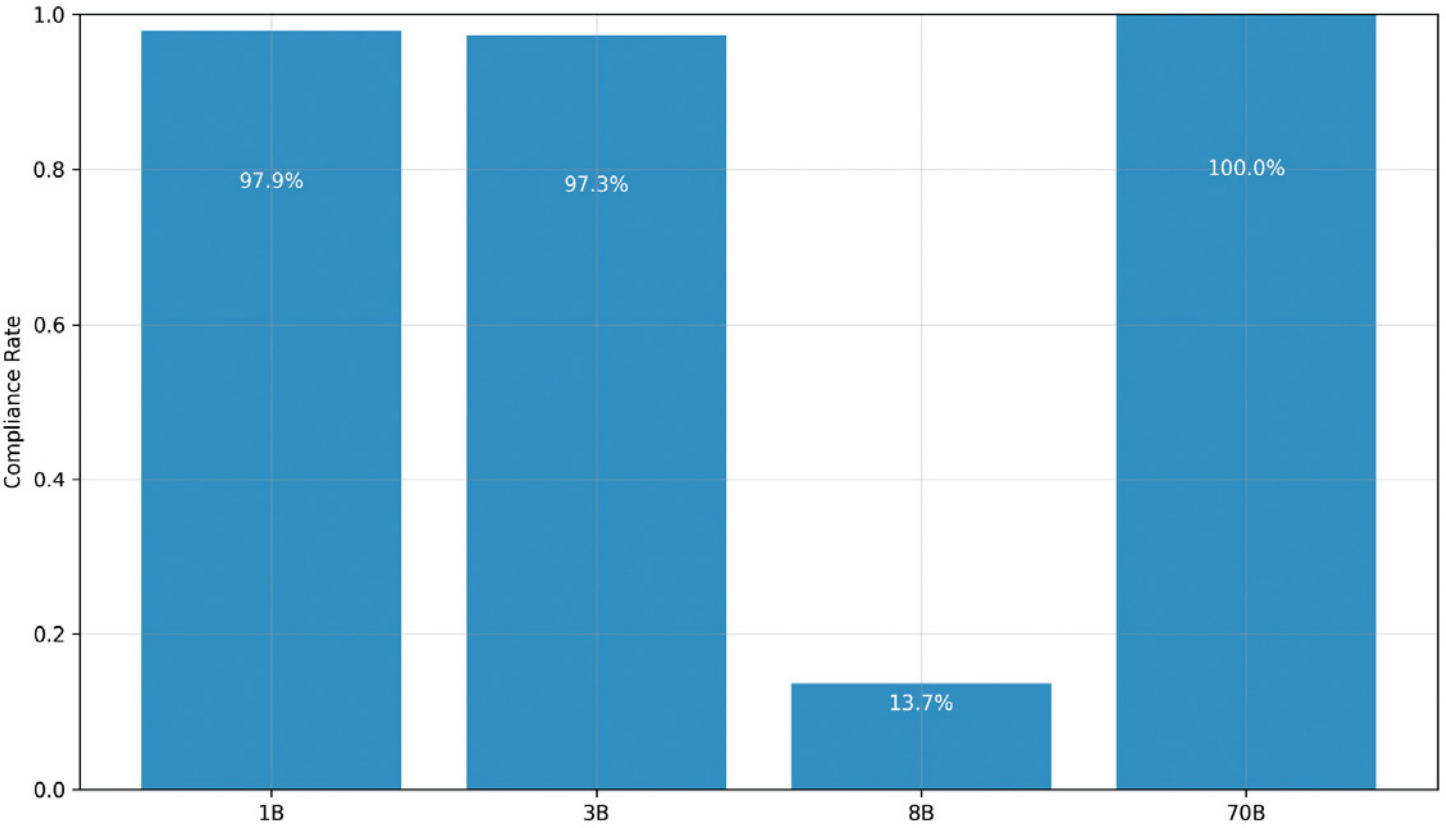
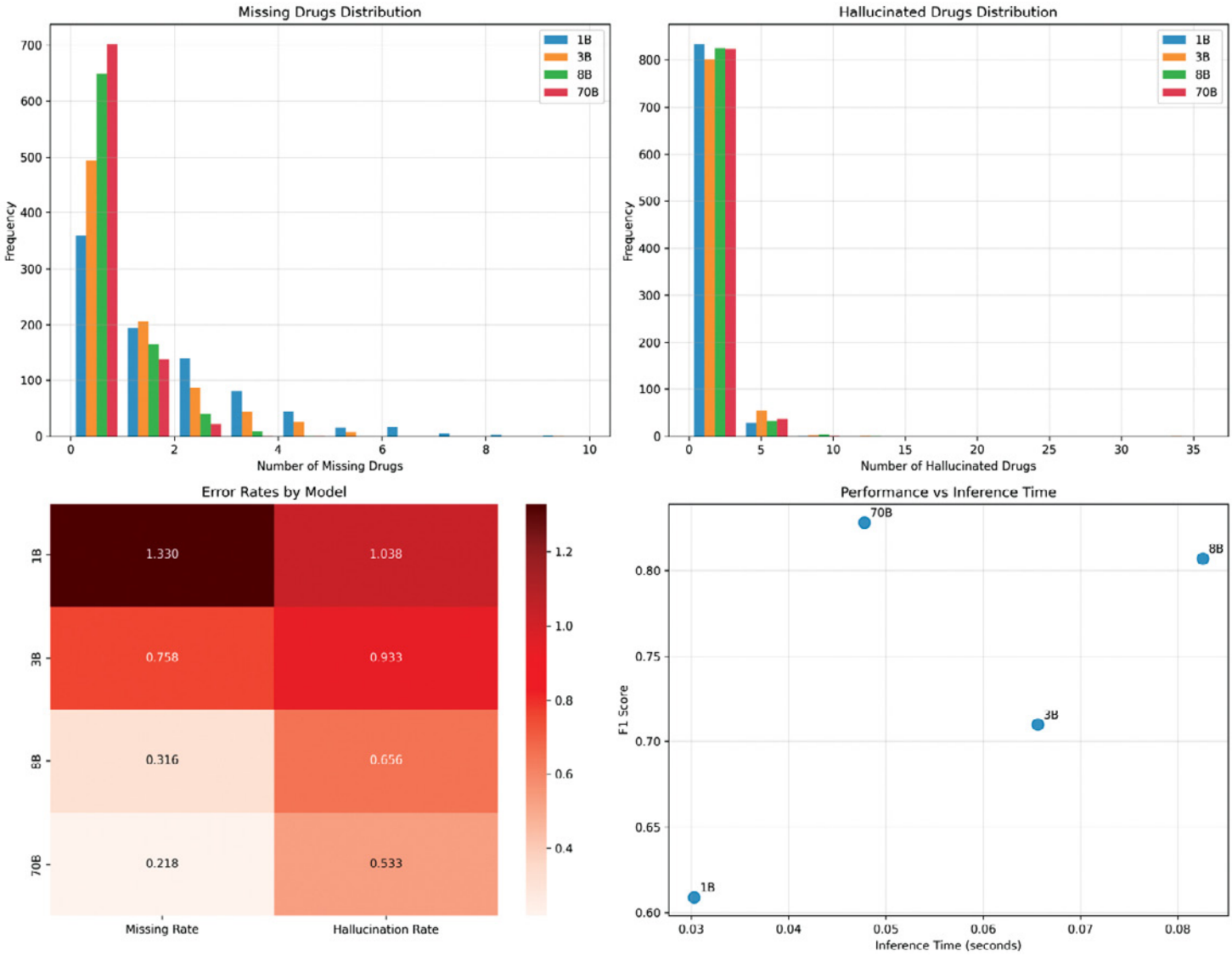


Figure 5. Error Analysis: Distributions of Per-Record Counts of Missing and Hallucinated Drugs (Top), Heatmap of Mean Counts per Record for Missing and Hallucinated Drugs (Bottom Left), and F1 Versus Mean Inference Time Per Record (Bottom Right)



than either the 8B or 3B model. This is likely due to highly compliant outputs that required fewer generated tokens. The mean per-case processing times ranked as follows: 1B (fastest), followed by 70B, 3B, and 8B (slowest).

Error profile

Both missing and hallucination rates declined with scale. Here, hallucinated drugs denote predicted drug entities that are absent from the expert-annotated gold standard for that record after normalization (i.e., false positives under a strict gold-standard comparison). The mean hallucinated-drug count per record decreased from 1.038 (1B) and 0.933 (3B) to 0.656 (8B) and 0.533 (70B) (Figure 5, bottom left). When false positives occurred, they tended to appear in small multi-drug bursts rather than as isolated single additions (Figure 5, top right), consistent with occasional expansion of regimen/protocol shorthand under a strict gold-standard comparison. Post-normalization, frequent misses were dominated by vincristine, cyclophosphamide, and asparaginase (Table 2).

Performance by text length

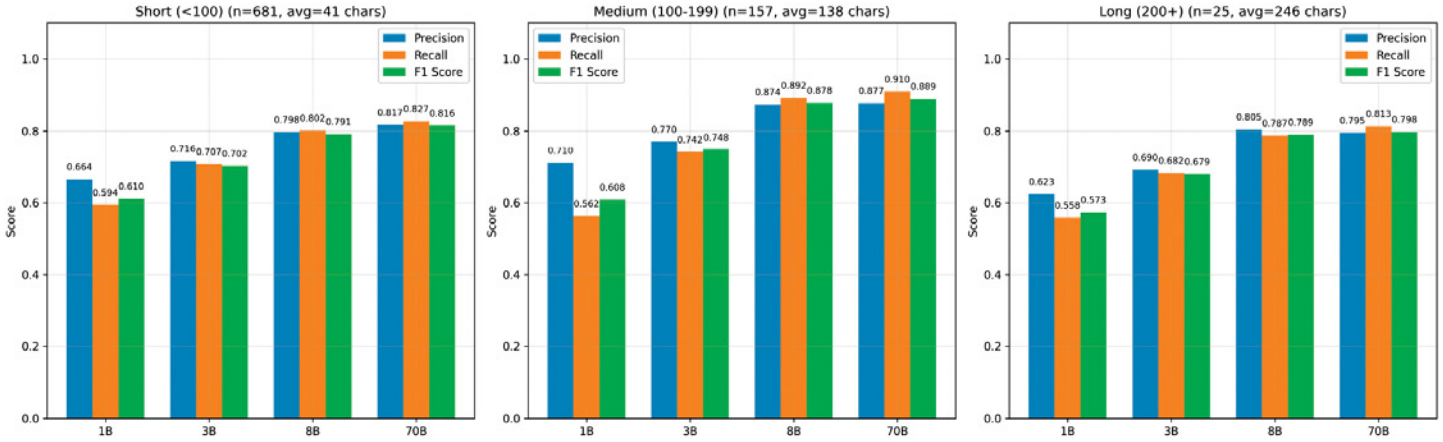
Scaling trends held across input lengths (Figure 6). Medium-length notes (100–199 characters; n=157) yielded the highest scores for 8B/70B (F1: 0.878/0.889), short notes (< 100; n=681) were uniformly lower, and long notes (200; n=25) showed mild declines relative to medium length while preserving the model ranking.

Discussion and Conclusion

This study demonstrates that a pipeline integrating constrained prompting and entity normalization provides reliable extraction of oncology drugs from clinical text. Performance improves significantly with model size, although the accuracy gains between the 8B and 70B models are modest compared to the substantial improvements observed when scaling from smaller (1B/3B) to mid-sized models. Robust normalization notably enhances performance, particularly benefiting smaller models, though even the largest (70B) model experiences meaningful improvements in precision and recall.

Table 2. Top 5 Missing and Hallucinated Drugs *After* Normalization by Model (Counts Across *N* =863 Cases)		
Model	*Top-5 Missing (Count)*	*Top-5 Hallucinated (Count)*
1B	vincristine (193), cyclophosphamide (163), etoposide (96), asparaginase (83), doxorubicin (67)	cytarabine (176), cisplatin (131), methotrexate (99), cabozantinib (43), asparaginase (37)
3B	cyclophosphamide (129), asparaginase (54), vincristine (50), daunorubicin (42), doxorubicin (34)	vincristine (68), doxorubicin (65), pegaspargase (36), cytarabine (34), ifosfamide (31)
8B	asparaginase (55), cyclophosphamide (29), daunorubicin (13), doxorubicin (13), vincristine (13)	pegaspargase (56), doxorubicin (44), vincristine (44), cyclophosphamide (38), bleomycin sulfate (31)
70B	asparaginase (55), cyclophosphamide (16), doxorubicin (10), temozolomide (8), busulfan mel-phalan (7)	pegaspargase (57), doxorubicin (39), bleomycin sulfate (32), vincristine (31), cyclophosphamide (28)

Figure 6. Performance by Text-Length Group (Short/Medium/Long) After Normalization



Post-hoc analysis suggests that many perceived "hallucinations" reflect difficulties in instruction adherence rather than genuine errors. For instance, models often listed individual constituent drugs when provided with regimen protocols, indicating limitations in instruction compliance rather than actual hallucinations. Consistent with this observation, increased model size corresponded to fewer omissions and extraneous drugs (Figure 4). Notably, the 8B model exhibited deviations from this trend, frequently repeating JSON objects and thereby negatively affecting its compliance metrics. In operational deployments, false positives could potentially be reduced via stronger prompt constraints and/or post-generation evidence checks requiring a supporting mention in the source text. We did not explore these prompt and filtering variants here to keep the scaling comparison controlled under a single fixed prompt and consistent post-processing.

Instruction compliance was identified as a crucial factor influencing both accuracy and processing efficiency. Models strictly adhering to JSON-only output protocols produced cleaner results and shorter latencies. Conversely, over-generation significantly increased generation time. The 70B model processed cases faster than the 8B and 3B models despite its size, benefiting from near-perfect compliance and minimal extraneous output. While the 1B model remained the fastest, its accuracy was notably lower. These insights highlight actionable optimization strategies-such as grammar or JSON-schema-constrained decoding, output gating, and truncation of non-JSON content to improve reliability and throughput, especially for mid-sized models.

Future work will investigate enhancing the pipeline through retrieval-augmented generation (RAG), specifically targeting drug and regimen knowledge. This approach will involve retrieving candidate aliases and regimen components from the normalization database during decoding and employing token-level logit masking or biasing to prioritize retrieved canonical generics (retrieval-augmented decoding). Additionally, formal assessment of RAG will be conducted through ablation studies evaluating retrieval effectiveness, varying the number of retrieved candidates, and examining latency trade-offs. Structured generation will also be expanded by explicitly incorporating a JSON schema within the prompt to constrain outputs, encompassing structured fields such as drug name, dosage, units, route of administration, frequency, and regimen details including component expansion and temporal anchors.

Several limitations should be considered in interpreting these findings. The extraction focused exclusively on drug names, without assessing detailed aspects such as dosing, timing, intent, or explicit regimen sequencing beyond downstream normalization. Furthermore, this study did not evaluate the impact of varying decoding and sampling hyperparameters (e.g., temperature, top-*p* / top-*k* settings, repetition penalty, max-token length), nor did it compare decoding strategies (beam search versus greedy). Alternative prompt formulations, including varying instructions or incorporating few-shot exemplars, also were not explored. Model comparison was restricted to the Llama family for a controlled scaling analysis; performance and error profiles may differ for domain-tuned biomedical

models, which we plan to evaluate in future work. Finally, evaluation was confined to a single corpus, highlighting the need for broader multi-institutional studies to confirm generalizability across diverse clinical note formats.

In summary, a constrained-prompt pipeline integrating robust normalization effectively extracts oncology treatment information from clinical text. Mid-to-large-scale LLMs demonstrate practical utility for clinical extraction tasks, with the 70B model offering the highest accuracy paired with predictable latency. However, the 8B model represents an optimal balance between accuracy and computational resources, with further potential improvements through schema-constrained and retrieval-augmented decoding. This study serves as a blueprint for deriving real-world evidence from cancer registries; enabling researchers and policymakers to better understand the nuances of cancer treatment and outcomes in diverse patient populations, ensuring that innovations are grounded in the realities of clinical practice.

References

1. Noone AM, Lund JL, Mariotto A, et al. Comparison of SEER Treatment Data With Medicare Claims. *Med Care* 2016;54(9):e55–e64.
2. Mallin K, Palis BE, Watroba N, et al. Completeness of American Cancer Registry Treatment Data. *J Registry Manag* 2013;40(2):59–66.
3. Jagsi R, Abrahamse P, Morrow M, et al. Under-ascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. *Cancer* 2012;118(2):333–341.
4. Enewold L, Parsons H, Zhao L, et al. Updated Overview of the SEER–Medicare Data: Enhanced Content and Applications. *Med Care* 2020;58(12 Suppl 2):S3–S13.
5. Grattafiori A, et al. The Llama 3 Herd of Models. *arXiv preprint* arXiv:2407.21783. 2024. Available from: **https://arxiv.org/abs/2407.21783**
6. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint* arXiv:2307.09288. 2023. Available from: **https://arxiv.org/abs/2307.09288**
7. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst* 2022;35:27730–27744.
8. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y, eds. *Proceedings of EMNLP 2022*. Abu Dhabi: ACL; 2022:1998–2022.
9. McInerney D, Young G, van de Meent JW, Wallace B. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In: Bouamor H, Pino J, Bali K, eds. *Findings of ACL: EMNLP 2023*. Singapore: ACL; 2023:8477–8494.
10. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf* 2019;42(1):99–111.
11. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner Ö. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020;27(1):3–12.
12. Miranda-Escalada A, Mehryary F, Luoma J, et al. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database (Oxford)*. 2023;2023:baad080. doi:10.1093/database/baad080
13. Keraghel I, Morbieu S, Nadif M. Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study. *arXiv preprint* arXiv:2401.10825. 2024. Available from: **https://arxiv.org/abs/2401.10825**
14. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems 33*. 2020:9459–9474.
15. Wang X, Wei J, Schuurmans D, et al. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. *arXiv preprint* arXiv:2203.11171. 2022.
16. Dong Q, Li L, Dai D, et al. A survey on in-context learning. *arXiv preprint* arXiv:2301.00234. 2023.
17. Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A survey on hallucination in large language models. *arXiv preprint* arXiv:2309.01219. 2023.
18. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation, and applications. *J Am Med Inform Assoc* 2010;17(5):507–513.
19. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17(3):229–236.
20. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267–D270.
21. Sohn S, Clark C, Halgrim SR, Liu H. MedXN: an open-source medication extraction and normalisation tool for clinical text. *J Am Med Inform Assoc* 2014;21(5):858–865. doi:10.1136/amiajnl-2013-002190
22. Almeida JR, Silva JF, Matos S, et al. A two-stage workflow to extract and harmonise drug mentions from clinical notes into observational databases. *J Biomed Inform* 2021;120:103849. doi:10.1016/j.jbi.2021.103849
23. Bodenreider O, Rodriguez LM. Analyzing U.S. prescription lists with RxNorm and the ATC/DDD Index. *AMIA Annu Symp Proc* 2014;2014:297–306.
24. HemOnc.org. Hematology Oncology Drug and Regimen Database. Accessed February 29, 2024. Available from: **https://hemonc.org**
25. Warner JL, Dymshyts D, Reich CG, et al. HemOnc: a standard vocabulary for chemother-apy regimen representation in the OMOP common data model. *J Biomed Inform* 2019;96:103239.
26. National Cancer Institute. Cancer Medications Enquiry Database (CanMED). SEER Oncology Toolbox. Version 1.26.1, 2025. Accessed July 10, 2024. Available from: **https://seer.cancer.gov/oncologytoolbox/canmed/**
27. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;46(D1):D1074–D1082.
28. National Library of Medicine. RxNorm. 2025. Accessed May 6, 2024. Available from: **https://www.nlm.nih.gov/research/umls/rxnorm**
29. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011;18(4):441–448.
30. National Cancer Institute. NCI Thesaurus (NCIt). 2025. Accessed July 10, 2024. Available from: **https://evs.nci.nih.gov**
31. National Cancer Institute. SEER*Rx Interactive Antineoplastic Drugs Database. 2025. Accessed July 10, 2024. Available from: **https://seer.cancer.gov/seertools/seerrx**
32. Clinical Trials Transformation Initiative. AACT Relational Database. 2025. Accessed July 10, 2024. Available from: **https://aact.ctti-clinicaltrials.org**
33. Tasneem A, Aberle L, Ananth H, et al. The database for Aggregate Analysis of Clin-icalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS One* 2012;7(3):e33677.
34. DOE–NCI MOSSAIC. *NCCR Treatment Database* (GitHub repository). 2025. Accessed September 26, 2025. Available from: **https://github.com/DOE-NCI-MOSSAIC/nccr-treatment-database**

How I Did It

CaRDO: A Tool for Empowering Cancer Registries to Publish Population-Level Cancer Statistics Online

James Retell, PhD[a]; Sean Francis, BDataSci[a]; James Rene, BMath[a]; Peter Baade, PhD[a,b]

***Abstract*:** Introduction: Population-based cancer registries are critical for monitoring cancer burden, informing policy, and guiding cancer control strategies. However, many registries face challenges in effectively communicating their data, due to limited resources and technical capacity. There is a growing need for accessible tools that support the visualization and dissemination of cancer statistics in engaging and interpretable formats. Methods: CaRDO (Cancer Registry Data Online) was developed by the Viertel Cancer Research Centre at Cancer Council Queensland as a free, secure, and user-friendly R package for creating interactive dashboards using population-level cancer data. Built on the R Shiny framework, CaRDO enables users with minimal programming experience to generate dashboards entirely offline. The tool incorporates standard epidemiological methods, including age-standardized rates, lifetime risk estimation, and segmented regression for trend analysis, while maintaining strong privacy safeguards. Results: CaRDO allows users to display cancer incidence and mortality data by sex, cancer type, year, age group, and measure (counts or rates), with visualizations of underlying temporal trends. All data processing occurs locally, and dashboards can be customized with plots that can be downloaded for use in reports or presentations. Example datasets and detailed guidance are provided to support implementation. The current version includes core dashboard elements, with flexibility for future expansion. Conclusions: CaRDO provides a timely and practical solution for enhancing the visibility and utility of cancer registry data. By removing financial and technical barriers, it supports under-resourced registries in fulfilling their essential role in cancer control through transparent, data-driven public health communication.

Key words: *Cancer registry, Cancer Registry Data Online (CaRDO)*

Introduction

Population-based cancer registries are essential for understanding the impact that cancer has at a population level. By systematically collecting standardized data, cancer registries provide a critical platform for tracking cancer incidence, survival, and mortality; monitoring trends over time; and identifying disparities in cancer indicators across population subgroups and geographical areas. They also provide a unique data-driven evidence base to support the development, implementation, and evaluation of cancer control efforts, from prevention and screening strategies to broader health system planning.[1,2]

The global cancer burden is rising rapidly. In 2022, there were an estimated 20 million new cancer diagnoses and 9.7 million cancer deaths. By 2050, annual cancer diagnoses are projected to reach 35 million, a 77% increase.[3] While outcomes have improved for some population groups and cancer types, significant disparities persist.[3] At the same time, academic research faces growing challenges, including funding shortfalls, public mistrust, pressure for translational research, and disconnects between evidence and policy.[4] In this context, it is more important than ever to effectively communicate the value of routine cancer surveillance through cancer registries.[4]

Making information derived from cancer registry data more accessible to policymakers, healthcare professionals, and the public is one way to raise the profile of cancer registries and highlight their value. Cancer registries should be viewed not just as data repositories, but as essential tools for guiding evidence-based decision making. Yet a major barrier is the lack of affordable, user-friendly tools for presenting cancer statistics in an engaging and accessible format. Many registries operate with limited resources, making it difficult to invest in proprietary software or specialized expertise. As a result, their ability to produce high-quality outputs that inform policy and engage stakeholders is often constrained.

Several software packages—such as Power BI, Tableau, and others—are purpose-built for data visualization and dashboard development. These tools offer powerful capabilities and are widely used across industry, including health analytics. However, they often require licensing costs and technical expertise, which can pose challenges for smaller or under-resourced cancer registries. While well-suited for building interactive dashboards, they are not specifically designed to meet the unique needs of population-based cancer registries aiming to communicate broad cancer trends to non-technical audiences.

In addition to these general-purpose platforms, there are more specialized tools designed for clinical and molecular cancer research.[5,6,7] However, tools focused on visualizing and communicating population-level cancer outcomes in accessible formats remain limited.

Florensa and colleagues[8] (2023) proposed a comprehensive workflow for visualizing cancer registry data using R Shiny dashboards. Their approach enhances data accessibility and transparency, supporting evidence-based

Address correspondence to: Professor Peter Baade, Cancer Council Queensland, Brisbane, Australia. Email: peterbaade@cancerqld.org.au
[a]Cancer Council Queensland, Brisbane, Australia; [b]Queensland University of Technology, Brisbane, Australia

decision-making. However, it relies on backend infrastructure—such as APIs and databases—to connect registry data to the dashboard application. This requirement may exceed the technical capacity of smaller or under-resourced registries.

To help bridge the gap in accessible online reporting tools for cancer registries, the Viertel Cancer Research Centre at Cancer Council Queensland has developed Cancer Registry Data Online (CaRDO). CaRDO is a free R package designed to support the creation of interactive dashboards using population-level cancer data. It was built specifically for registries with limited technical capacity—requiring no programming skills and ensuring that sensitive data remain offline.

Methods

CaRDO (Cancer Registry Data Online) was developed by the Viertel Cancer Research Centre at Cancer Council Queensland to provide a free, secure, and user-friendly solution for visualizing population-level cancer statistics. The tool was designed to support cancer registries with limited technical capacity, enabling them to create interactive dashboards without requiring programming expertise or online access to sensitive data.

Technical Framework

CaRDO is implemented as an R package using the R Shiny[9] framework from the R programming language.[10] It allows users to build interactive web-based dashboards that display cancer statistics in a clear and engaging format. The package is freely available and only requires installation of R Studio, which is also free. All data processing occurs locally on the user's computer, ensuring that sensitive information remains secure and offline.

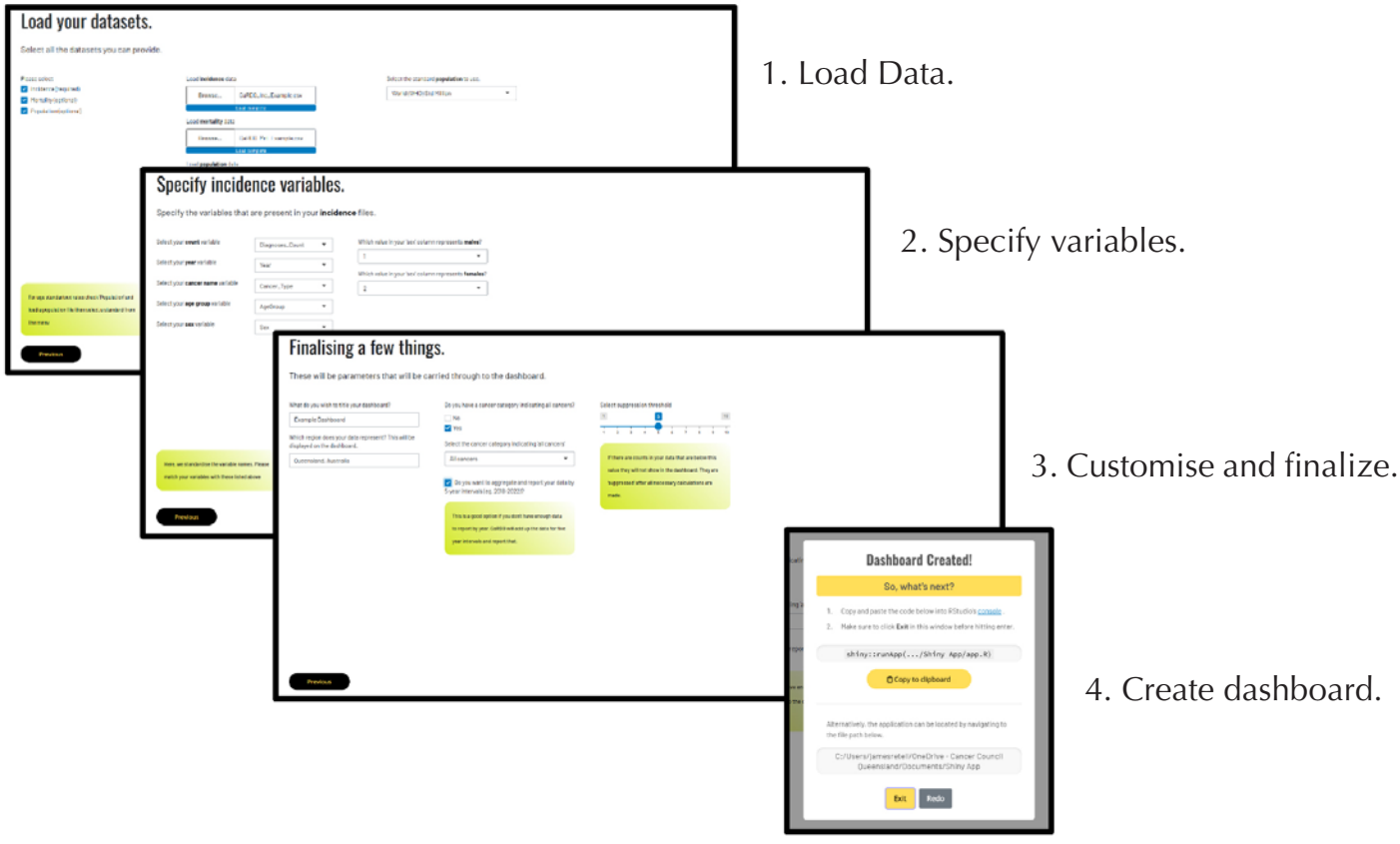
Broadly, CaRDO works by standardizing the user's data structure and variable names, allowing the prebuilt R Shiny application to perform the required calculations on the supplied data. The package, including the software and its documentation, is only distributed via GitHub, a free online code-hosting platform, which allows users to install CaRDO directly from within R. The full source code is publicly available at **https://github.com/** for readers and users wishing to explore the technical side of CaRDO in more detail. Importantly, GitHub is not used to store or process user data; all analyses are performed locally on the user's computer.

User Workflow

CaRDO was designed to be intuitive and accessible to users with minimal or no experience in R programming. The dashboard creation process is guided entirely through a graphical user interface, with no coding required. The workflow involves four main steps (see Figure 1): (1) loading data into R Studio; (2) specifying variables of interest; (3) customizing dashboard details; and (4) generating the final dashboard on the local computer.

Some initial data preparation is required prior to these steps, and detailed guidance is provided in the

Figure 1. Steps Involved in Generating a Statistics Dashboard Using CaRDO. For further details, see *The CaRDO Handbook*, Chapter 4, CaRDO Example



accompanying **The CaRDO Handbook**. The handbook also outlines best practices for data formatting, aggregation, and privacy protection. In addition, CaRDO has built-in error handling to identify common user issues—such as mis-specified variable names or invalid data types—and provides clear feedback to users to help them identify and correct these errors.

The content of the CaRDO dashboard depends on the scope of the data available. For example, the simplest dashboard users can create provides a display of cancer incidence count data. This requires details about cancer type, counts (number of diagnoses), diagnosis year, age at diagnosis (in 5-year age groups), and sex. This can then be expanded to calculate incidence rates, if the corresponding population data are available. If mortality data are available, they can be incorporated into the dashboard using the current version of CaRDO.

Statistical Methods

When counts and populations were provided, age-standardized rates were calculated using the direct method, applying age-specific rates to the world standard population across 5-year age groups. Lifetime risk estimates were derived using cumulative risk formulas, assuming constant age-specific rates up to age 85.

Trends in cancer incidence and mortality were modeled using segmented regression techniques. Breakpoints were identified using structural change detection methods, and piecewise models comprised of linear fits were applied to each segment.

Sex-specific cancers were identified algorithmically by assessing the presence of sex-coded data across cancer types. Where data are available for only one sex, the cancer type is classified as sex-specific, and so the person-specific population count is not used as the denominator.

Privacy and Data Governance

CaRDO was developed with strong privacy safeguards to ensure compliance with data governance standards. All data used in CaRDO are stored and analyzed locally; no data are transmitted or stored online. For incidence and mortality dashboards, only aggregated data are required, eliminating the need for unit record data. Even when sensitive data are used, CaRDO functions analogously to other secure data management software, importing and processing data entirely within the RStudio environment.

To further protect privacy, CaRDO includes several built-in features including (a) 5-year aggregation options to reduce the risk of identifying individuals in registries with small case numbers, (b) user-defined suppression thresholds to automatically exclude small cell counts from display, and (c) offline operation to ensure data never leave the user's device.

Development and Testing

The software was built using the following R packages: shiny software package for R,[9] tidyverse,[11] bslib,[12] shinyWidgets,[13] plotly,[14] markdown,[15] strucchange,[16] and segmented.[17]

CaRDO was developed iteratively with input from internal and external stakeholders, including cancer researchers, cancer registrars, clinical coders, and data custodians, both within Australia and internationally. Feedback was collected through individual discussions, either in person or through videoconference and/or email, with the intent to refine usability and ensure clarity of the documentation and relevance to privacy standards and registry practices. Before deployment, CaRDO was tested using a range of internal and external datasets, including both real-world registry data and purpose-built fictitious datasets designed to test data security and suitability for sparse datasets. As this version of CaRDO represents the first iteration, ongoing user feedback is actively encouraged to inform future development and refinement.

Publishing the Dashboard

The focus of CaRDO is to produce a web-based dashboard that is stored on the local computer. CaRDO does not automatically publish the data online, thus providing the user with greater security while using CaRDO and more flexibility in terms of how they want to publish the dashboard online. There are several options available for publishing this dashboard, and their respective strengths and limitations have been summarized elsewhere.[18]

Example Datasets

To help users implement and test CaRDO, example datasets containing fictional incidence, mortality, and population data are provided to demonstrate the required data structure.

Results

An example of the dashboard that can be generated by CaRDO is provided in Figure 2. Depending on the data provided, the dashboard can display population-level cancer data by sex, cancer type, year, age group, and measure (rates or counts) for both incidence and mortality, and details of the underlying temporal trends.

All data displayed in the dashboard, as well as the figures themselves, are available for users to download, enabling straightforward incorporation into reports, presentations, or other materials.

Discussion

The development of CaRDO aligns with a growing recognition of the vital role cancer registries play in public health. Cancer registries are not merely data repositories—they provide a critical foundation essential to appropriately understand the impact of cancer within a population, guide policy, and drive improvements in cancer outcomes. Yet despite their importance, registries often operate under resource constraints and so face challenges in communicating their value to stakeholders. CaRDO directly addresses this gap by providing a free, user-friendly tool that enables registries to transform their confidential complex data into accessible, public-facing interactive dashboards.

By removing financial and technical barriers, CaRDO provides the potential to empower under-resourced registries to enhance the visibility and impact of their data. This

Figure 2. Example "Diagnosis" Page Dashboard Created Using CaRDO with Statistics Based on Fictional Data

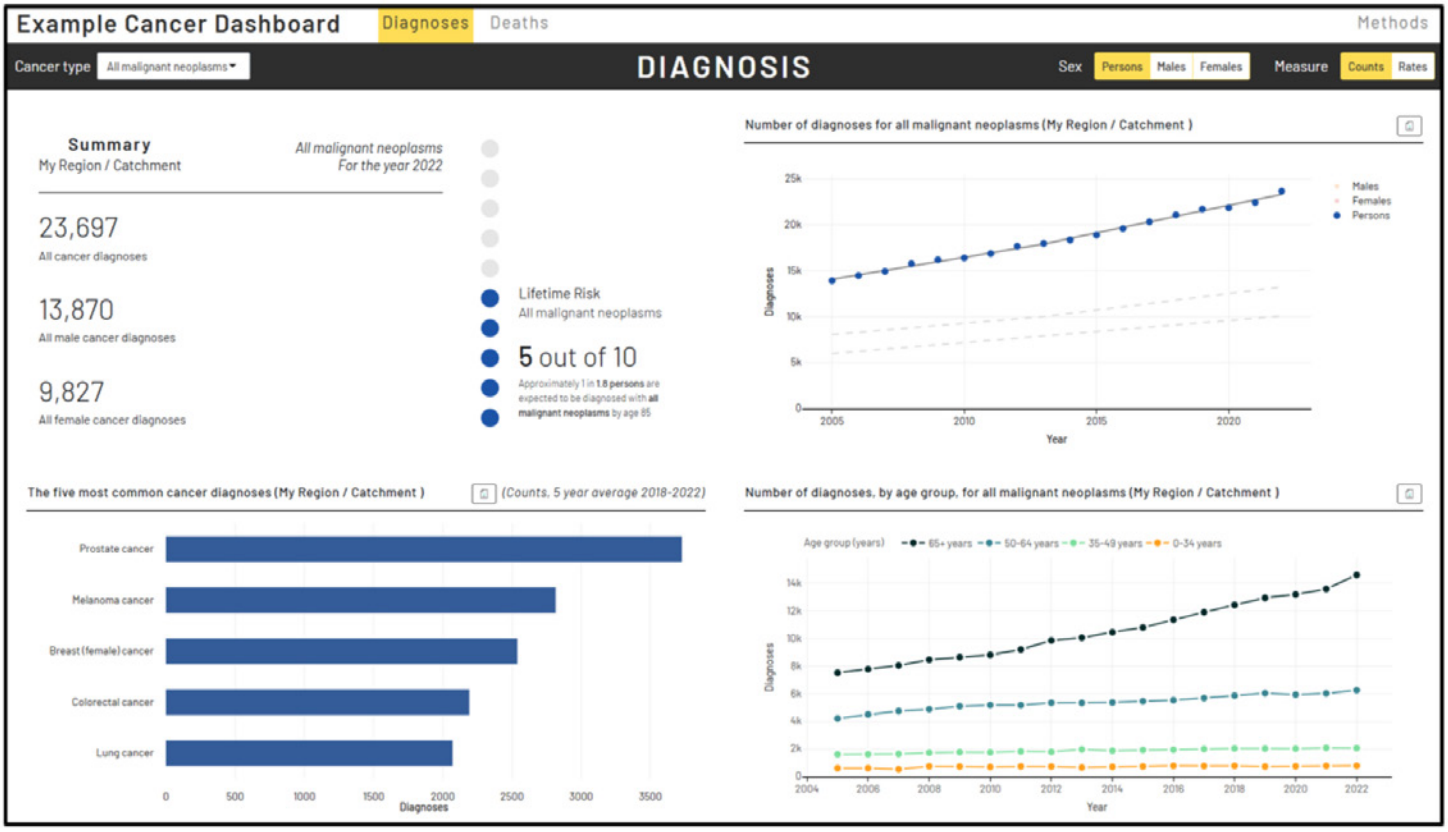
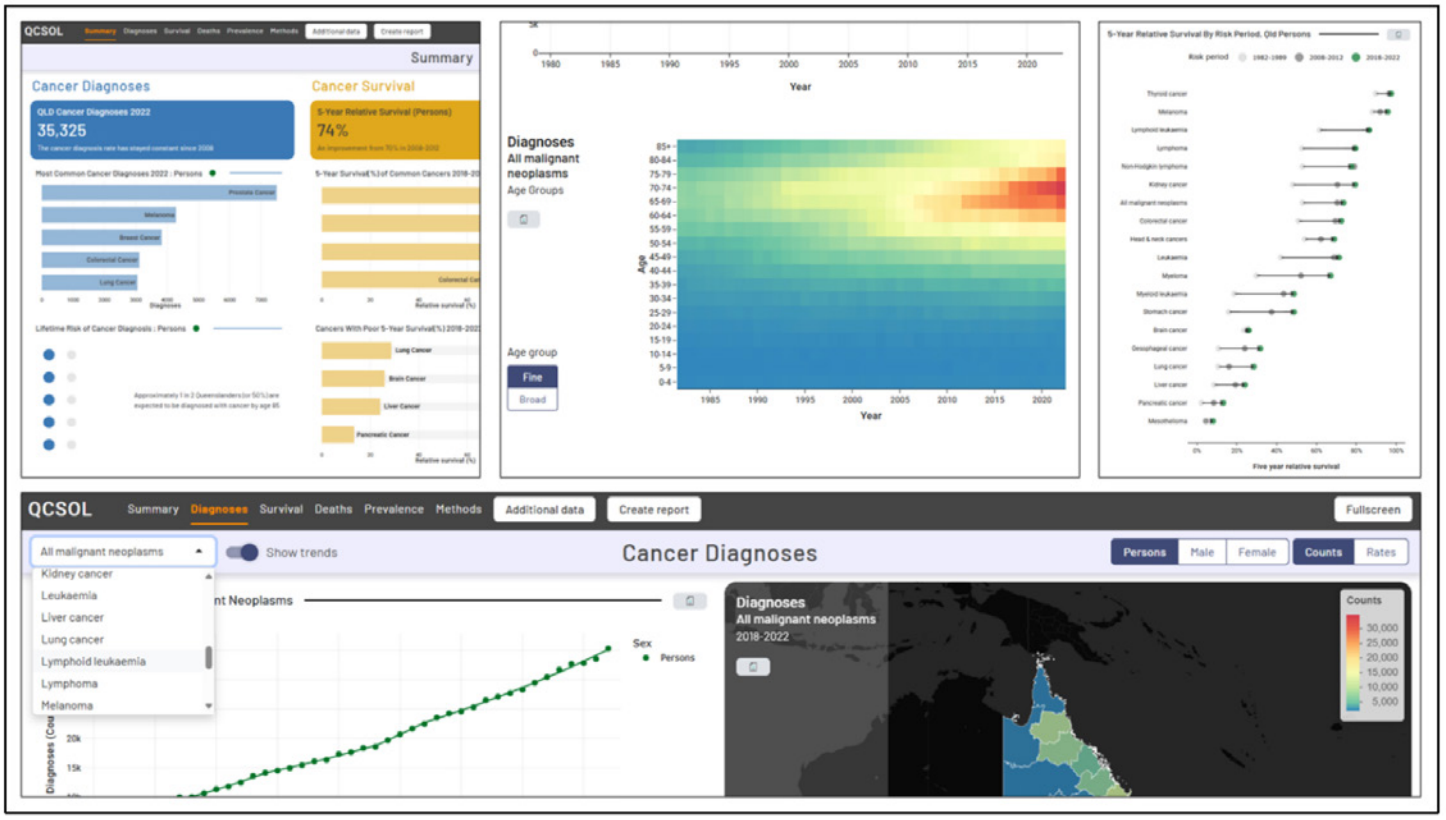


Figure 3. Collage of Dashboards from Queensland Cancer Statistics Online (QCSOL), Demonstrating Potential for Future Expansion of CaRDO



democratization of data visualization supports broader efforts to champion cancer registries and raise awareness of their contributions to cancer control. In doing so, CaRDO helps registries better engage with policymakers, healthcare professionals, and the public—key audiences whose support is essential for sustaining and expanding registry operations.

The current version of CaRDO has been designed to include the core elements that might typically be expected in a standard online dashboard for a population-based cancer registry. Utilizing the inherent flexibility of R Shiny, there are many avenues for expansion, such as including various measures of survival and more details about trends and age-specific patterns, along with many options for presenting these statistics visually. The Queensland Cancer Statistics Online (QCSOL)[19] provides one example of the possibilities that could be included with further development of the underlying CaRDO package (see Figure 3).

The CaRDO software and associated materials are freely available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). This license permits users to copy, redistribute, and adapt the materials for non-commercial purposes, provided appropriate credit is given and any derivative works are shared under the same license. This open licensing model supports collaborative development and equitable access, particularly for low-resource cancer registries. Organizations seeking customizations or enhancements beyond the base product are encouraged to contact the developers to discuss potential syntax development on a cost-recovery basis.

Conclusion

CaRDO provides a secure, accessible, and adaptable solution for visualizing population-level cancer data. Its intuitive interface and open-source design make it especially valuable for registries with limited resources, enabling them to produce interactive dashboards without programming expertise or compromising data privacy. Built on a flexible R Shiny framework, CaRDO supports a range of statistical outputs and can be expanded to meet evolving registry needs. As cancer burden continues to rise and the need for transparent, evidence-based public health responses grows, CaRDO provides a timely and impactful tool to help cancer registries fulfill their essential role in cancer control.

While CaRDO is designed to be user-friendly, users are welcome to contact statistics@cancerqld.org.au for general enquiries or support. If your organization requires customizations or enhancements to the base product, please contact us to discuss further. Feedback from users is actively encouraged, as understanding how CaRDO is adopted and used is critical to demonstrating its impact and supporting future development.

References

1. Parkin, D.M. The role of cancer registries in cancer control. *Int J Clin Oncol*. 2008;13(2):102–11.
2. Van Eycken, L.J., et al. Future of population-based cancer registries: A global perspective—A survey of population-based cancer registries. *Int J Cancer.* 2025;157(8):1566–1576.
3. Bray, F., et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–263.
4. Piret, G., et al. A call to action to address escalating global threats to academic research. *Innovation (Camb),* 2025.;6(4):100758.
5. Petrov, I. and A. Alexeyenko. EviCor: Interactive Web Platform for Exploration of Molecular Features and Response to Anti-cancer Drugs. *J Mol Biol.* 2022;434(11):167528.
6. Deng, M., et al. Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. *BMC Bioinform.* 2016;17:72.
7. Yang, I.S., et al. ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. BMC Genom. 2016;17(1):631.
8. Florensa, D., et al. Exploring cancer incidence, risk factors, and mortality in the Lleida region: interactive, open-source R Shiny application for cancer data analysis. *JMIR Cancer.* 2023;9:e44695.
9. Chang, W., et al. *shiny: Web Application Framework for R.* 2024.
10. Team, R.C., *R: A Language and Environment for Statistical Computing.* 2025.
11. Wickham, H., et al. Welcome to the tidyverse. *J Open Source Softw.* 2019;4(43):1686.
12. Sievert, C., J. Cheng, and G. Aden-Buie. *bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'.* 2024.
13. Perrier, V., F. Meyer, and D. Granjon. *shinyWidgets: Custom Inputs Widgets for shiny.* 2024.
14. Sievert, C. *Interactive Web-Based Data Visualization with R, plotly, and shiny.* Chapman and Hall/CRC. 2020.
15. Xie, Y., J.J. Allaire, and J. Horner. *markdown: Render Markdown with 'commonmark'.* 2024.
16. Zeileis, A.. Implementing a Class of Structural Change Tests: An Econometric Computing Approach. Computational Statistics & Data Analysis, 2006;50(11):2987–3008.
17. Muggeo, V.M.R., et al.. Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling,* 2014;14:293–313.
18. Pereira, R. *R shiny Deployment: 5 Options for Individuals and Enterprises*. 2024 [cited 2025 04/09/2025]; Available from: **https://www.appsilon.com/post/r-shiny-deployment**.
19. Queensland, C.C. Queensland Cancer Statistics Online. 2025 September 2024 [cited 2025 September 4]; Available from: **https://cancerqld.org.au/research/explore-our-projects/queensland-cancer-statistics-online/**.

How I Did It

The Population Cancer Assessment and Surveillance Engine (PopCASE): An Emerging Population Cancer Data Platform

Johnie Rose, MD, PhD[a]; Fatima Hussain, MPH[a]; Satya S. Sahoo, PhD[a]; Kalle Lyytinen[b]; Sunah Song, PhD[c]; Harry J. Menegay, PhD[c]; Robert Lanese, MS[c]; Akhil S.G.[b]; Lihua Liu, PhD[d]; Jennifer Tsui, PhD[d]; Mark Beno, MSM[c]; Todd Burus, PhD[e]; Siran M. Koroukian, PhD[a]

Abstract: Spatially linking cancer registry records to information describing the communities where patients live can enhance the utility of registry data. Here, we describe the ongoing development of an application, the Population Cancer Assessment and Surveillance Engine (PopCASE), that will allow central cancer registries or their designees to set up a population cancer data platform linking a registry's data with community data. PopCASE links registry records by patient residential address to demographic, socioeconomic, household, transportation, healthcare access, screening, risk factor, and other community metrics. Users can query the data by demographic, geographic, or disease criteria via a user interface and receive results, including case counts, age-adjusted incidence and mortality, stage at diagnosis, time-to-treatment, and dozens of community measures stratified by county, census place ("municipality"), zip code tabulation area (ZCTA), and census tract. PopCASE may also be used to create patient-level linked research data sets. We anticipate two likely implementation scenarios: 1) a cancer center where a research team with appropriate institutional review board (IRB) approval uses an instance of PopCASE for their catchment-level population cancer research in a secure data environment, and 2) a central cancer registry that hosts their own secure instance of PopCASE.

Key words: *cancer, registry, informatics, social drivers of health, community*

Introduction

Cancer registry data represents the lifeblood of population cancer research and cancer control efforts. The value of linking cancer registry data to place-based contextual data is increasingly recognized. In 2024, the National Cancer Institute (NCI) released a version of the Surveillance, Epidemiology, and End Results (SEER) data linked to certain characteristics (socioeconomic status, rurality, and persistent poverty) of de-identified census tracts.[1] Prior to this, a handful of solutions emerged, packaging aggregated cancer incidence or mortality data (typically county-level) with community metrics from US Census data and elsewhere.[2-4] One notable dashboard featuring a rich collection of community metrics is Cancer InFocus,[4] developed at the University of Kentucky Markey Cancer Center, and adopted at 52 sites nationwide as of September 2025. Cancer InFocus allows a cancer center or other entity to easily implement a public-facing dashboard drawing data from publicly available sources, including site-specific, county-level cancer incidence or mortality rates.

While dashboards using public data expand the user base of cancer data to those with general questions about the distribution of cancer burden, these tools are not ideal for research. Nor are they well-suited for the type of sub-county-level surveillance and community outreach and engagement (COE) planning needed for localized cancer control efforts.[5,6] Researchers often need patient-level data, which commonly would be linked to community-level indicators (e.g., median income) through a laborious merging process. Public health and COE professionals may also need information on very specific groups of cancer patients, living in communities smaller than counties, to effectively target their efforts. For this reason, our team at Case Comprehensive Cancer Center developed a prototype application called the Ohio Cancer Assessment and Surveillance Engine (OH-CASE). Under institutional review board (IRB) approvals (Case Western Reserve University Protocol IRB-2016-1752 and Ohio Department of Health Protocol 2017-50), we linked records from the Ohio Cancer Incidence Surveillance System (OCISS) by address at diagnosis to county and sub-county-level social and economic indicators from the US Census and other sources.

Building on the success of the OH-CASE prototype, and with NCI funding, our team is incorporating numerous additional data sources and generalizing the model into a piece of software—the Population Cancer Assessment and Surveillance Engine (PopCASE)—to allow other central cancer registries or their approved data users to implement their own OH-CASE-like platform using patient-level cancer data linked with multifaceted community information.

Here, we provide an overview of PopCASE, report current project progress, and discuss implementation considerations.

Methods

PopCASE consists of a PostgreSQL relational database management system (DBMS), a user interface with support for querying and viewing the data, and a "controller" layer

[a]Case Western Reserve University School of Medicine, Cleveland, OH. [b]Case Western Reserve University Weatherhead School of Management, Cleveland, OH. [c]Cleveland Institute for Computational Biology, Case Western Reserve University School of Medicine, Cleveland, OH. [d]Keck School of Medicine of University of Southern California, Los Angeles, CA. [e]University of Kentucky Markey Cancer Center, Lexington, KY.

Address correspondence to Johnie Rose, MD, PhD 10900 Euclid Ave. LC-4889 Cleveland, OH 44106.

where any calculations needed to generate results in real time occur. The PopCASE DBMS is populated using an extract transform load (ETL) workflow with individual database tables. A key design principle of PopCASE is that all input data sources must be available in standardized formats for all parts of the US. North American Association of Central Cancer Registries (NAACCR) formatted patient-level cancer registry data serves as the backbone of the database, and any user implementing an instance of PopCASE must have access to registry data with IRB approval.

Figure 1 provides a summary of the data sources included in PopCASE. Below, we describe each community-level data source and its role in PopCASE.

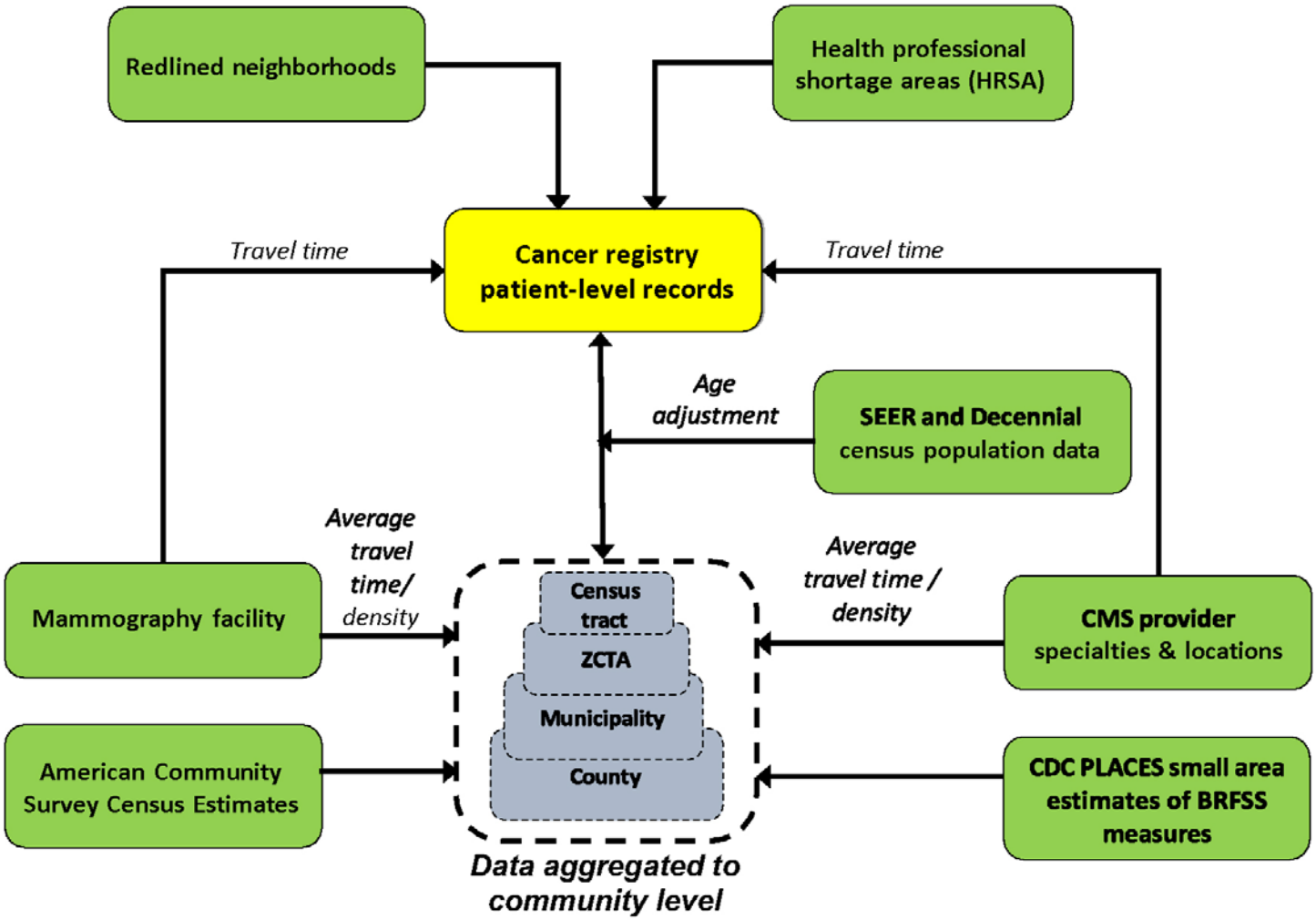
SEER and Decennial census population data[7,8]

PopCASE contains race/ethnicity-specific tables of age- and sex-stratified population counts for each county, census place ("municipality"), zip code tabulation area (ZCTA), and census tract in the region of interest (in our case, Ohio) at different time points. These counts are used to age-adjust incidence and mortality rates. Performing age-adjustment in real time allows PopCASE to generate age-adjusted incidence and mortality rates and confidence intervals for any combination of disease-, age-, race/ethnicity-, or location-based search criteria. Census tract and county population estimates are those produced by Woods & Poole Economics, Inc. with support from NCI.[7] Census place and ZCTA population estimates are derived from 2010 and 2020 decennial census data;[8] counts for intercensal years are estimated by linear interpolation, while 2020 data is used for post-2020 years.

Five-year U.S. Census American Community Survey (ACS) data[9]

These data provide contextual understanding of the demographic makeup, economic circumstances, housing environment, household composition, residential stability, educational attainment, and transportation resources of communities at the county, municipality, ZCTA, and census tract levels.

Figure 1. Conceptual Diagram of PopCASE Data Sources and Their Linkages



CDC = Centers for Disease Control and Prevention, CMS = Center for Medicare and Medicaid Services, HRSA = Health Resources and Services Administration, SEER = Surveillance Epidemiology and End Results, ZCTA = ZIP code tabulation area.

CDC PLACES data

The Centers for Disease Control and Prevention (CDC) provides small area estimates of Behavioral Risk Factor Surveillance System (BRFSS) measures.[10] The database contains estimates of the proportion of residents of each county, municipality, ZCTA, and census tract who are current on screening for breast, cervical, and colorectal cancers; as well as prevalence estimates for cancer-related risk factors, including current smoking, binge drinking, and physical inactivity, and for certain comorbidities and healthcare access indicators.

Doctors and clinicians national downloadable file

This resource from the Centers for Medicare and Medicaid Services (CMS) provides information on clinical provider practice street address and specialty.[11] This data is used to estimate average travel distances to multiple provider types from the population-weighted centroid of each census block group (a proxy for patient address) and census tract. County-level provider-to-resident ratios are also calculated.

Food and Drug Administration (FDA) mammogram facility registry

This registry contains the street addresses of all certified US mammography facilities.[12]

Redlined neighborhoods

In over 200 U.S. cities, 1930's federal mortgage lending guidelines rated neighborhoods where Black Americans and members of other minority groups resided as high risk, shutting off their access to mortgage credit.[13] PopCASE creates indicators for each patient of residence in a former redlined (highest risk) neighborhood based on historical Home Owner Loan Corporation (HOLC) lending maps.[14]

Health Professional Shortage Areas

PopCASE creates patient-level indicators of residence in a designated primary care health professional shortage areas (HPSA).[15]

Filters

The PopCASE interface is used to build queries based on combinations of patient demographics (sex, age groups, race/ethnicity), county, diagnosis date, cancer type(s), and stage.

Aggregation

Query results may be output at the level of county, municipality, ZCTA, or census tract (one row of data per location). With appropriate IRB permissions, PopCASE can output a patient-level data set. While all PopCASE visual output is tabular, data can be exported as a csv file for further analysis or visualization with external tools.

Measures

Users may select from dozens of measures calculated from the patient- and community-level PopCASE data. For geographically aggregated results, these include disease-focused measures such as case/death counts, age-adjusted incidence/mortality, time to treatment, stage distribution, or average Gleason score (prostate cancer). National incidence and mortality estimates from the Cancer in North America (CINA) public use data set[16] are provided for comparison. Each row of output will also contain user-selected community measures including travel time to various healthcare services, screening and health services utilization, economic measures, household characteristics, transportation access, etc.

Stratification

Geographically aggregated results can be stratified by numerous demographic, insurance status, or disease characteristics (including breast cancer receptor status).

Results

The PopCASE project has completed its first year. With the database built, the next steps are to enhance the prototype user interface and build the controller logic for completing age adjustment and other real time calculations. Here, we provide some statistics describing OH-CASE, the Ohio implementation of PopCASE residing in the Case Western Reserve University Secure Research Environment.

OH-CASE contains population data for approximately 11.8 million Ohioans living in 88 counties, 1,204 census places/municipalities, 1,197 ZCTAs, and 2,952 census tracts. Community data from the sources shown in Figure 1 is available for each of these geographic units. The database contains registry records for 880,463 Ohioans with cancer diagnosed from 2010 through 2022. The total storage requirement for the database is approximately 18 gigabytes.

Our team's version of PopCASE (OH-CASE) utilizes OCISS data, but OCISS data will not be packaged with the PopCASE application.

Discussion

We have described the design and current state of PopCASE, a population cancer analytics platform designed to spatially link numerous community-level data sources to identified patient-level registry data. The final stage of the project entails building a "software container" which allows the application to be installed in various computing environments. We will package the location-specific portions of each constituent public data set with the container used by any particular adoptee to minimize setup burden.

The technical work of developing PopCASE, while substantial, is relatively straightforward; but the governance considerations are critical and will likely vary by adopting site. Registry data of adopting sites will not be shared outside the environment where PopCASE is implemented (including with our development team). PopCASE will be distributed essentially as a wrapper of curated community data that links to the adoptee's own patient-level registry data during setup. This approach offers tremendous practical advantages relative to manual processes that involve acquiring, preparing, and integrating these data sources with registry data using SEER*Prep/SEER*Stat or other tools. It is also important to note that any spatial linkage activities involving cancer registry data depend on accurate

geocoding. Given the value of such linkage (exemplified by, but not the sole domain of, PopCASE), adequate financial support and appropriate tools for registries' geocoding activities represent an important advocacy priority for the cancer registry community.

We foresee two likely implementation settings: 1) cancer centers where a research team with appropriate IRB approval uses an instance of PopCASE for population cancer research in a secure data environment, and 2) a central cancer registry hosting their own secure instance of PopCASE. In the latter scenario, PopCASE could be used by registry or public health staff to fulfill data requests or perform advanced surveillance on a region's cancer burden. In addition, registries would have the option of developing a portal for vetted, credentialed users in the cancer control community (such as cancer center COE teams) to securely access geographically aggregated data.

Acknowledgments

Cancer incidence data used in these analyses were obtained from the Ohio Cancer Incidence Surveillance System (OCISS), Ohio Department of Health (ODH), a cancer registry partially supported by the National Program of Cancer Registries at the Centers for Disease Control and Prevention (CDC) through Cooperative Agreement Number NU58DP007097. Use of these data does not imply that ODH or CDC agrees or disagrees with the analyses, interpretations, or conclusions in this report (or publication/presentation). Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA284198. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

The authors report no conflicts of interest, financial or otherwise.

References

1. Bhattacharya M, Cronin KA, Farrigan TL, Kennedy AE, Yu M, Srinivasan S. Description of census-tract-level social determinants of health in cancer surveillance data. *J Natl Cancer Inst Monogr.* 2024;2024(65):152–61. doi: 10.1093/jncimonographs/lgae027.
2. Sylvester Comprehensive Cancer C. Scan 360 2020 [cited 2024 May 20]. Available from: **https://www.scan360.com/**
3. National Cancer I. State Cancer Profiles 2023 [cited 2024 May 3]. Available from: **http://statecancerprofiles.cancer.gov**.
4. Burus JT, Park L, McAfee CR, Wilhite NP, Hull PC. Cancer InFocus: Tools for cancer center catchment area geographic data collection and visualization. *Cancer Epidemiol Biomarkers Prev.* 2023;32(7):OF1–OF5. doi: 10.1158/1055-9965.EPI-22-1319.
5. DelNero PF, Buller ID, Jones RR, Tatalovich Z, Vanderpool RC, Ciolino HP, Croyle RT. A National Map of NCI-Designated Cancer Center Catchment Areas on the 50th Anniversary of the Cancer Centers Program. *Cancer Epidemiol Biomarkers Prev.* 2022;31(5):965–71. doi: 10.1158/1055-9965.EPI-21-1230.
6. Paskett ED, Hiatt RA. Catchment Areas and Community Outreach and Engagement: The New Mandate for NCI-Designated Cancer Centers. *Cancer Epidemiol Biomarkers Prev.* 2018;27(5):517–9. doi: 10.1158/1055-9965.EPI-17-1050.
7. National Cancer Institute. U.S. Population Data 2025 [cited 2025 June 30, 2025]. Available from: **https://seer.cancer.gov/data-software/uspopulations.html**.
8. U.S. Census Bureau. Decennial Census of Population and Housing Datasets 2024 [cited 2025 June 30]. Available from: **https://www.census.gov/programs-surveys/decennial-census/data/datasets.html**.
9. United States Census Bureau. About the American Community Survey 2023 [cited 2024 May 1]. Available from: **https://www.census.gov/programs-surveys/acs/about.html**.
10. Centers for Disease C, Prevention. PLACES: Local Data for Better Health 2024 [cited 2025 June 30]. Available from: **https://www.cdc.gov/places/index.html**.
11. U.S. Center for Medicare and Medicaid Services. Doctors and Clinicians National Downloadable File 2025 [cited 2025 September 26]. Available from: **https://data.cms.gov/provider-data/dataset/mj5m-pzi6**.
12. Food and Drug Administration (FDA). Mammogram Facilities 2024 [cited 2024 April 14]. Available from: **https://www.fda.gov/radiation-emitting-products/mammography-information-patients/search-certified-facility**.
13. Bailey ZD, Feldman JM, Bassett MT. How structural racism works—racist policies as a root cause of U.S. racial health inequities. *N Engl J Med.* 2021;384(8):768–73. doi: 10.1056/NEJMms2025396.
14. Nelson RK, Winling L, Marciano R, Connoly N. Mapping Inequality: Redlining in New Deal America 2022 [cited 2024 May 20]. Available from: **https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58&text=downloads**.
15. U.S. Health Resources and Services Administration (HRSA). HPSA Finder 2022 [cited 2022 October 1]. Available from: **https://data.hrsa.gov/tools/shortage-area/hpsa-find**.
16. North American Association of Cancer Registries. CiNA Public Use Data Set 2025 [cited 2025 June 30]. Available from: **https://www.naaccr.org/cina-public-use-data-set/**.

How I Did It

Modernizing Registry Operations: Implementing File*Pro Automation with Groovy Scripting to Streamline Data Processing in Missouri

Nishant Jain, PhD, MS, MHA, FAMIA[a,b]; Deborah Carey, ODS-C[a,b]; Lucinda Ham, RHIA, ODS-C[a,b]; Jeffrey Steffens, ODS-C[a,b]; Joseline Hernández, MPH[c]; Iris Zachary, PhD [a,b,d], MS, FAMIA, ODS-C[a,b,c]

Abstract: Background and objective: Processing out-of-state (OOS) cancer registry files has historically required time-consuming manual work and programmer intervention. At the Missouri Cancer Registry and Research Center (MCR), File*Pro was integrated into our process at MCR with state-specific groovy configuration files to streamline file preparation, reduce turnaround time, and empower oncology data specialists (ODS) to work independently. This report evaluates the effectiveness of this approach and shares adaptable templates for replication by other registries. Methods: Custom groovy scripts were developed to align OOS files with PrepPlus requirements. These scripts automated default value population, corrected state-specific fields (e.g., county codes), and provided reusable templates. Evaluation metrics included pre- and post-implementation processing times, staff independence, and QA dashboard indicators. Results: From January to April 2025, 180 cases across 9 OOS files were processed with minimal programmer support. County code corrections applied to 3,476 cases from 31 files using the same workflow. ODS staff independently processed up to 13 OOS files in 4 months, saving ~5 days per file by avoiding programmer delays. QA metrics showed a 5-day improvement in turnaround times. Conclusion: File*Pro with groovy configurations offer a scalable, adaptable solution for registries. By reducing reliance on programmers, empowering staff, and accelerating workflows, this approach enhances efficiency and can be replicated in other central registries with minimal training.

Key words: *cancer registry, File*Pro, groovy scripts, automation, workflow efficiency*

Introduction

Central cancer registries (CCRs) serve as the backbone of cancer surveillance in the United States.[1,2] Their core responsibilities include collecting, consolidating, and standardizing cancer incidence data from hospitals, laboratories, and other reporting sources.[1–3] This data supports cancer research, informs public health programs, and guides policy decisions at state and national levels.[8,10] However, CCRs also manage complex datasets that extend beyond their geographic boundaries. One such initiative involves the interstate exchange of cancer data: particularly when patients are diagnosed or treated outside their state of residence. As data timeliness becomes an increasing priority in national surveillance efforts,[10] central registries must adopt more automated and responsive approaches to handling file variability and exchange protocols, including out-of-state submissions.

These out-of-state (OOS) files often arrive in varying formats, contain non-standard fields, and must be harmonized with the receiving registry's specific requirements. Without streamlined processes, this can lead to inefficiencies, delays, and increased dependency on specialized technical staff. It is often necessary to adjust out-of-state (OOS) exchange files to account for Missouri-specific state layouts, non-standard fields, and mismatched county codes. Historically, this process has been heavily manual or dependent on programmers, which has caused bottlenecks and delays. One had to either work with Excel to solve these issues or wait for a programmer to create SAS or SQL code. This limited the speed at which registry data pre-processing contributions could be made and placed strain on IT support staff.

Nearly half of cancer registrars say they need additional training in data analysis, and recent workload statistics show that they spend more than an hour on simple abstracts and almost two hours on difficult abstracts.[1–3] File*Pro is a software utility developed and maintained by the National Cancer Institute (NCI)-SEER to assist central cancer registries with data preparation, recoding, and validation prior to submission through PrepPlus and the NAACCR Call for Data workflows.[7] Even with the availability of programs like File*Pro, underutilization of technology is still a significant problem that frequently leads to inefficiencies and lost value in a variety of businesses. Cross-sector enterprise analyses have claimed that organizations, including healthcare and public-sector entities, use only about 30–40% of the capabilities of their digital tools, leading to significant waste of investment and staff time, even when capable software has been available.[4–6] In a way, this project also addresses the underutilization gap by leveraging already

[a]Missouri Cancer Registry and Research Center, University of Missouri, Columbia, MO. [b]College of Health Sciences, Department of Public Health, University of Missouri, Columbia, MO. [c]Missouri Department of Health and Senior Services, Jefferson City, MO. [d]MU Institute for Data Science and Informatics, University of Missouri, Columbia, MO.

Address correspondence to Nishat Jain, Missouri Cancer Registry and Research Center, University of Missouri, Columbia, MO, 65211.

existing tools, such as File*Pro and groovy configuration files, to reduce processing time, empower Oncology Data Specialists (ODS), and scale workflows for OOS file preparation. The immediate objective was to replace manual and code-heavy operations or processes with automated workflows for OOS file processing using File*Pro + groovy configuration files. The future vision is to build a sustainable, scalable solution that empowers ODS staff to execute processes independently, reducing reliance on programmers while enhancing efficiency, consistent with national public health surveillance strategy.[10]

Methods

The Missouri Cancer Registry and Research Center (MCR) is supported in part by a cooperative agreement between the Centers for Disease Control and Prevention (CDC) and the Missouri Department of Health and Senior Services and a surveillance contract between DHSS and the University of Missouri. Figure 2 depicts the prior workflow for processing OOS files, which relied heavily on programmer availability for configuration and execution. In the new process, following file receipt, ODS staff load the incoming NAACCR XML into File*Pro, select the appropriate Groovy configuration, and execute automated recoding prior to PrepPlus import (Figures 1 and 2). Prior to broader rollout, the groovy-based workflow was piloted with a subset of ODS staff and refined based on feedback to ensure alignment with existing file preparation tasks. Initial training materials were developed in-house, and follow-up support was offered to ensure familiarity with File*Pro's scripting interface. Custom groovy scripts were created for each OOS file type. The groovy scripts apply standardized defaults, enforce state-specific logic, and prepare the file for downstream QA without requiring programmer intervention (Table 1). These scripts populated default values required by PrepPlus software, corrected county codes to Missouri standards, and automated repetitive tasks. Older workflows involved manual changes by ODS for 5 min, 30 min and 120 min in some cases or 1–5 day wait times for programmers. The revised workflow can be summarized as: OOS file → Groovy script → File*Pro → PrepPlus. (Figures 1 and 2).

Two time-based metrics were tracked for this study:

- Processing time, referring to the actual staff effort per file, and
- Programmer wait time, defined as the elapsed delay between a request and a programmer's response.

To estimate potential time savings, we modeled 2 common scenarios based on an annual registry workload of 120 OOS files:

- A simple scenario requiring 5 minutes of ODS-led processing per file.
- A complex scenario requiring 30 minutes per file.

Efficiency gains were attributed to reduced manual effort and elimination of programmer delays. Two pre-automation delay models were considered:

- Scenario A: A per-file 5-day (40-hour) programmer wait.
- Scenario B: A batch delay, where five files incur a shared 1-day wait.

These scenarios provide a framework within which to evaluate both individual and systemic time savings and are summarized in Table 2 and Figure 2.

Results

Between January and April 30, 2025, the Missouri Cancer Registry (MCR) processed 180 cases from 9 out-of-state (OOS) files using the new File*Pro and groovy-based automation approach. Projected over a full year, this translates to approximately 540 cases from 27 OOS files. In addition, the system was successfully adapted for other use cases, such as county code corrections (3,476 cases from 31 files) and Veterans Affairs (VA) data, using the same groovy configuration templates. The automation workflow—from file ingestion to completion—is depicted in Figure 1 and the impact of the chances is shown in Figure 2. Staff feedback indicated that ODS staff independently processed up to 13 OOS files within a 3-month span, reporting a consistent 5-day reduction per file in processing time. Over 85% of OOS files were fully processed by ODS staff without requiring programmer intervention.

Quality assurance (QA) dashboard metrics, as well as feedback from ODS staff, confirmed a 5-day improvement in turnaround time and a reduction in EDIT error rates post-implementation.

The remaining 15% of files that still required programmer involvement were limited to edge cases. These could be NAACCR metafile updates, uncommon site-specific staging logic, or schema changes not yet integrated into reusable groovy scripts, especially when it exceeded the File*Pro GUI capabilities and/or required advanced groovy scripting.

To assess time savings, we modeled both individual file and batch processing scenarios, accounting for manual processing time (5–30 minutes per file) and programmer wait times (1–5 days). These scenarios are detailed in Table 2 and Figure 2. Even under conservative assumptions (for files that needed 5–30 mins prep time), projected gains include elimination of 968 to 4,808 hours annually in programmer queue delays—freeing technical staff to focus on higher-impact tasks. Together, these results demonstrate that the implementation not only reduced dependency on programmers but also empowered ODS staff, improved QA timelines, and expanded automation potential across multiple file types.

Discussion and Conclusion

This MCR Project primarily helped with staff empowerment and efficiency. Training ODS staff to use File*Pro scripts reduced dependence on limited IT staff.[1,2] This approach also improved staff autonomy and ownership, which became evident in their confidence to manage OOS files independently.[2,3] Despite an increasing number of OOS files received, the introduction of groovy-enabled

Figure 1. File*Pro Interface with groovy Advanced Editor

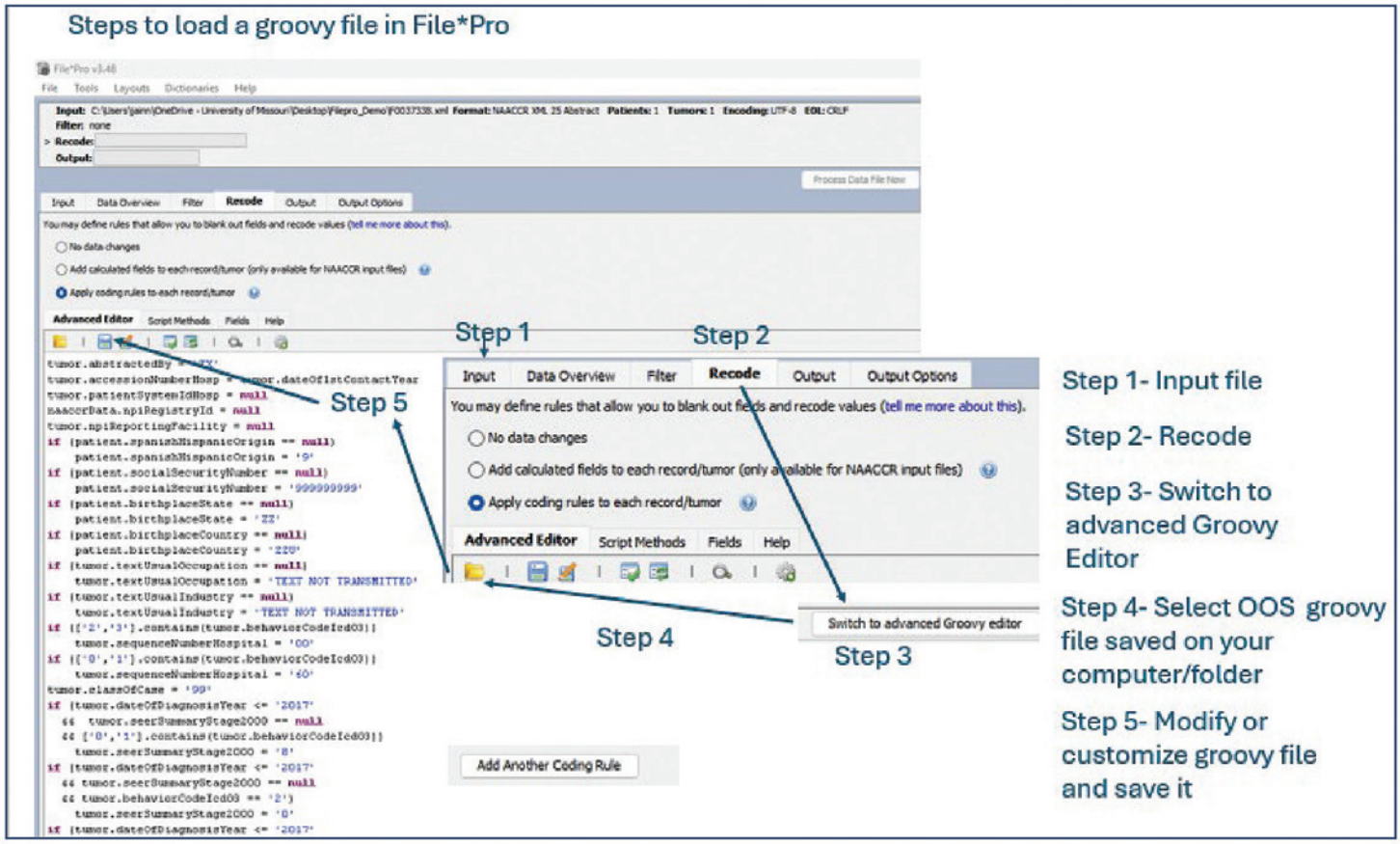
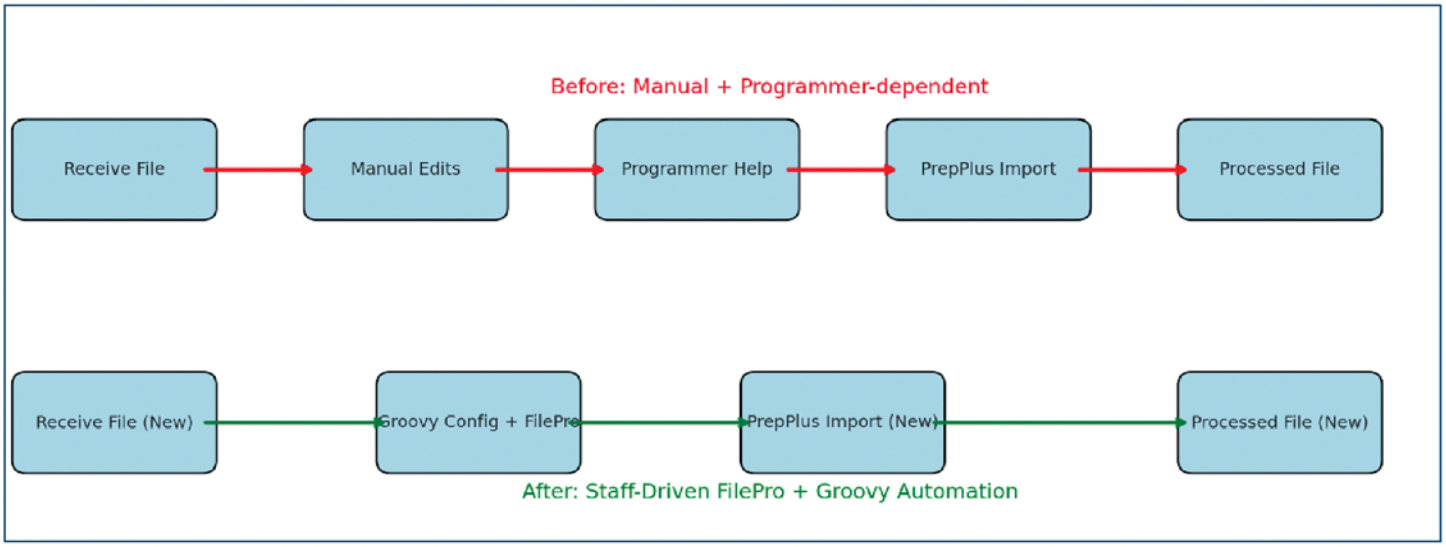


Figure 2. Workflow Comparison (Before Versus After Automation)



File*Pro processing has allowed the registry to maintain or improve throughput without expanding programmer capacity and allowing programmers to focus on higher-level integration tasks and enhancements.[1,3] Moreover, the approach is scalable, and the templates are portable. Any registry with access to File*Pro can adapt our groovy configurations with minor adjustments.[7,9] This is especially beneficial for registries with constrained staffing or programming resources.[1,3] By championing utilization of technology, this project highlights a practical response to the underuse of powerful tools.[4–6] By removing technical barriers, we democratized access to File*Pro's automation potential and eliminated hidden inefficiencies from our pipeline and time savings (Table 2).

We developed a groovy-based preprocessing and recode pipeline including a suite of groovy scripts within File*Pro to automate preprocessing of heterogeneous registry data files prior to PrepPlus processing and

quality assurance (Figure 1). These scripts standardize core fields by applying NAACCR-conformant unknown codes, normalizing identifiers (e.g., clearing National Provider Identifier (NPI), resetting Medical Record Number (MRN), and enforcing consistent staging logic by diagnosis year (e.g., assigning SEER Summary Stage 2000 for cases diagnosed before 2018 and Summary Stage 2018 thereafter). Modules also manage state-specific adjustments such as county reassignment for out-of-state cases, targeted recodes for death certificate–only (DCO) submissions, and normalization of incoming Veterans Affairs data (Table 1). Collectively, these processes ensure

Table 1. File*Pro groovy Scripts: Purpose and Example Rules for Registry Operations

Script / Module	Purpose	Example Rules only (check groovy scripts for comprehensive set of rules)
OOS24.groovy	Normalize out-of-state abstracts prior to PrepPlus/QA	• Assign NAACCR-conformant unknowns (e.g., ZZ/ZZU for state/country, 99999 postal code, 998 county, "TEXT NOT TRANSMITTED" for narratives). • Clear NPIs and standardize MRN placeholders. • Set classOfCase=99 for OOS records. • Derive sequenceNumberHospital from ICD-O-3 behavior (2/3→00; 0/1→60). • Apply staging by diagnosis year (≤2017→SEER Summary Stage 2000; ≥2018→Summary Stage 2018). • Null Collaborative Stage fields for DX ≥2016.
county-fix.groovy	Enforce in-state semantics for Missouri	• If patient.addrCurrentState ≠ MO → set patient.countyCurrent = 998. • If tumor.addrAtDxState ≠ MO → set tumor.countyAtDx = 998. • If patient.countyCurrent is null → backfill from tumor.countyAtDx.
DCO-fixes.groovy	Standardize Death Certificate Only (DCO) cases	• Set naaccrRecordVersion = 250. • For primarySite=C739 (thyroid), set schemaDiscriminator1=1. • For histologies 9732 or 9591 (hematopoietic/lymphoid), set schemaDiscriminator1=9. • Assign unknown grades (8) for sites C421 and C779.
VA.groovy	Normalize VA submissions to mirror NAACCR defaults	• Apply sequence number rules by behavior (2/3→00; 0/1→60). • Default Missouri-specific tobacco/alcohol items when null. • Normalize treatment narratives to "TEXT NOT TRANSMITTED".

Table 2. Comparison of File Processing Time and Projected Savings Before and After Implementation of File*Pro + Groovy Automation.

Table 2 summarizes three modeled scenarios comparing pre- and post-implementation time estimates for ODS-led file processing. Two dimensions of efficiency gains are shown: (1) prep time reduction for automated processing (from 5, 30, or 120 minutes per file down to ~1 minute), and (2) elimination of programmer wait time under two assumptions: a worst-case per-file delay (5 days = 40 hours for one file), and a conservative batch model (5-day delay = 40 hours for 5 files). All values are based on a workload of 120 files per year. These estimates align with staff-reported feedback and are intended to illustrate potential/projections savings across varying registry workflows based on actual reported savings for four months.

Scenario	*Before (hours)*	*After (hours)*	*Estimated Saved (hours)*	*Does estimate include programmer wait time? (Y/N/ NA=Not Applicable)*
Basic file processing time savings excluding programmer wait times				
Prep + QA only (5 min)	10	2	8	NA
Prep + QA only (30 min)	60	2	58	NA
Prep + QA only (120 min)	240	2	238	NA
Projected scenarios including programmer wait times in savings				
Per-File (Full 5-day Delay, 5-min prep)	4,810	2	4,808	Yes
Batch Model (1-Day Wait), 5-min prep	970	2	968	Yes
Per-File (Full 5-day Delay, 30-min prep)	4,860	2	4,858	Yes
Batch Model (1-Day Wait, 30-min prep)	1,020	2	1,018	Yes

that mixed-source inputs are transformed into edit-tolerant, registry-ready NAACCR XML files with consistent coding and minimized manual intervention. Customizable OOS and County groovy files and codes are available via GitHub.[9]

The Missouri implementation is designed to be portable, and other registries can adapt the scripts by updating state identifiers, unknown value codes, and policy-specific rules (e.g., county coding, class-of-case defaults, staging thresholds) (Figure 1, Step 5).[9] More advanced customization may involve extending site- or histology-specific recodes or aligning record versioning with local EDITS metafiles.[8] By parameterizing these elements in a simple configuration layer, registries can tailor the same automation framework to their own operational requirements while preserving the efficiency gains demonstrated in Missouri.

Limitations and Implications for Other Registries

This study's findings should be considered in light of a few limitations. The reported time savings and workflow enhancements were based on staff-reported estimates and qualitative feedback. This may introduce variability in precision. Initial setup time for groovy scripts was not included in per-file savings estimates. File*Pro training time was also excluded from our quantitative models. Although the initial learning curve for File*Pro scripting is modest, adoption may still require training and periodic support, especially in smaller teams. While the run-time for most files is under 1 minute once scripts are created, building and testing each configuration involves a one-time effort, which may vary depending on file complexity and registry-specific rules. Although feedback from MCR staff suggests a manageable learning curve, individual comfort levels and prior technical experience can influence the efficiency gains during early adoption phases.

Programmer wait times (1-day versus 5-day scenarios) were modeled to provide conservative and worst-case estimates but may not apply uniformly across all registries or use cases.

Despite these constraints, the broader implications are compelling. This approach offers a practical, scalable framework for modernizing registry workflows using low-code tools.[5,10] The scripting infrastructure is reusable across other file types-such as Veterans Affairs data, county corrections, and death certificate–only submissions—making it a valuable asset for future registry use cases.[7–9] By enabling frontline staff to take ownership of preprocessing tasks, we not only reduced turnaround time but also freed programmers to focus on higher-impact system development and integration efforts.[1–3] Overall, the Missouri experience demonstrates how scripting-enabled automation, when paired with thoughtful training and modular design, can yield rapid and sustainable improvements in cancer registry operations.[10]

Acknowledgments

Supported in part by Cooperative Agreement NU58DP007130-04 from the Centers for Disease Control and Prevention (CDC) and the Missouri Department of Health and Senior Services (DHSS).

Conflict of Interest Statement

The authors declare no conflicts of interest, financial or nonfinancial, related to the content of this manuscript.

Ethics/Consent Statement

This project involved process analysis, quality improvement, and methodological demonstration. No identifiable patient information was included in groovy scripts, and therefore informed consent was not required. All data handling complied with NAACCR data standards and state registry confidentiality policies.

References

1. Chapman SA, Hailer L, Miller J. Workload and Staffing Study of Hospital Cancer Registries. *J Registry Manag.* 2025;52(1):16–21.
2. Merriman KW, Broome RG, De Las Pozas G, Landvogt LD, Qi Y, Keating J. Evolution of the Cancer Registrar in the Era of Informatics. *JCO Clin Cancer Inform.* 2021;5:272–278. doi:10.1200/CCI.20.00123
3. UCSF–NCRA. *Final Workload Report 2024*. Accessed December 29, 2025. University of California, San Francisco; 2024.
4. WalkMe Ltd. *The State of Digital Adoption 2024*. Accessed December 29, 2025. **https://www.walkme.com/the-state-of-digital-adoption-2024/**
5. Bughin, J, LaBerge, L, Mellbye, A. *The Case for Digital Reinvention*. McKinsey 2017. Accessed December 29, 2025. **https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-case-for-digital-reinvention#/**
6. Greengard S. Tackling the High Cost of Unused Software. *CIO Insight.* November 18, 2014. Accessed December 29, 2025. **https://www.cioinsight.com/blogs/tackling-the-high-cost-of-unused-software/**
7. File*Pro Software Version 3.48 released June 9, 2025, **https://seer.cancer.gov/tools/filepro/**
8. NAACCR. Call for Data Submission Standards and Edits Metafile. North American Association of Central Cancer Registries; 2025. Available at: **https://www.naaccr.org**
9. Reusable Groovy configurations & Groovy scripts developed for the Missouri Cancer Registry (MCR) to automate File*Pro processing, streamline out-of-state (OOS) file preparation, apply county and DCO corrections, and improve registry workflow efficiency **https://github.com/nishrj/MCRFile-ProGroovy.git**
10. Richards CL, Iademarco MF, Anderson TC. A new strategy for public health surveillance at CDC: improving national surveillance activities and outcomes. *Public Health Rep.* 2014;129(6):472–476. doi:10.1177/003335491412900603

How I Did It

Addressing the Challenge of Geospatial Accuracy in Cancer Registration: The Louisiana Tumor Registry Experience

Christina Lefante, MPH, ODS-C[a]; Lauren Maniscalco, MPH[a]; Xiao-Cheng Wu MD, MPH, ODS-C[a]

***Abstract*:** Accurate geospatial data is foundational to cancer research, supporting efforts in cancer control and prevention programs, exposure analysis, and understanding of the socio-economic determinants of health.[1–5] As a population-based cancer registry, the Louisiana Tumor Registry (LTR) supplies data to national cancer datasets via North American Association of Central Cancer Registries (NAACCR), Centers for Disease Control (CDC), and Surveillance, Epidemiology, and End Results Program (SEER), while also publishing geographically-detailed cancer statistics for the state. Beyond these national contributions, LTR data has been instrumental in a wide range of cancer research applications, including but not limited to environmental exposure analysis.[6] Recognizing the need for cancer statistics at the community level, the Louisiana State Legislature enacted R.S. 40:1105.10 in 2018, requiring an annual cancer incidence report by census tract.[7] As a SEER registry, LTR has utilized the SEER Data Management System (SEER*DMS) since 2009, integrating the NAACCR-supported and NCI-funded geocoder, Geocodio API. While individual registries and jurisdictions will possess varying resources in obtaining complete and accurate geospatial data, all NAACCR registries have access to Geocodio. This report summarizes the strategies and practices LTR has implemented to achieve high quality geospatial data and provides guidelines that other registries can implement to meet this critical objective.

Key words: *geospatial, Louisiana Tumor Registry*

Methods

To support precise geospatial mapping of cancer incidence, the Louisiana Tumor Registry (LTR) has implemented a comprehensive workflow focused on improving the accuracy of address information at diagnosis. The precision of GIS mapping is directly dependent on the quality of the geographic coordinates, which in turn rely on clean and complete address data. LTR has established a set of best practices to determine the most accurate address at diagnosis and subsequently assign GIS coordinates. The main elements of this process are the geocoding workflow, in conjunction with address normalization and standardization, PO box resolution, and quality control checks.

Geocoding Workflow

The geocoding workflow includes an initial pass through the NAACCR-supported geocoder, Geocodio API, where geocoding is attempted on all cases with an address at diagnosis listed on the consolidated record. An address is uploaded to the Geocodio API and the matched address, along with the associated Census tract and GIS coordinates, are returned. When the submitted registry address matches the returned Geocodio address on all four input items (street address, city, state, zip code), LTR accepts the results and appends the geocoding/census data in the registry database. Cases that fail or return low confidence matches (e.g., any one of the four input items fails to match) are subjected to correction and reprocessing through a secondary, manual review, and then the address is sent back through the Geocodio API. It is common for this process to be repeated with complex cases and when necessary, alternative geocoding platforms (e.g., ArcGIS or Google Maps API) are used to resolve discrepancies. A final validation occurs when geocoded outputs are reviewed for consistency with known geographic boundaries and census tract assignments.

Address Normalization and Standardization

In any given year, LTR identifies a subset of cases with typographical errors, formatting inconsistencies, or incomplete address entities. These issues are typically flagged during the initial pass through the Geocodio API. To correct them, LTR employs a combination of automated and manual processes, which can be further classified into pre- and post-process standardization. During automated standardization, address data is first processed using USPS address validation tools and commercial reference databases to standardize formatting and correct common errors (e.g., missing street suffixes, invalid zip codes). Cases that remain unrecognizable undergo manual review and are reviewed by trained staff, who cross-reference hospital records, pathology reports, and other internal sources to resolve discrepancies.

PO Box Resolution

Approximately 5% of cases in Louisiana will have only a post office (PO) box reported as address at diagnosis, which should not be geocoded, since the coordinates would be of the post office's physical location and not a patient's residence. To address this challenge, LTR has developed a protocol that includes supplemental data sources, manual investigation, and documentation. Alternate addresses are

[a]Louisiana Tumor Registry, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA 70112.

Address correspondence to Christina Lefante, Louisiana Tumor Registry, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA 70112.

retrieved from supplemental data sources, such as medical records, billing information, or fee for service public databases. When alternate addresses are unavailable, staff may consult regional service maps or contact reporting facilities to determine a plausible residential location. Finally, all PO box resolutions are documented to ensure transparency and reproducibility.

Quality Control Checks

LTR conducts annual quality control activities to identify problematic addresses that are not standardized during the first run through the Geocodio API (about 2% of cases). These issues are typically due to typographical errors or data entry inconsistencies. LTR corrects problematic addresses prior to final geocoding. These quality control measures include flagging problematic addresses, setting a correction threshold, and leaving an audit trail. Cases are prioritized based on their impact on census tract-level mapping and legislative reporting requirements, while addresses that fail geocoding due to missing components, invalid characters, or ambiguous formatting are flagged for review. All corrections are logged, and geocoding success rates are monitored to inform future improvements.

Discussion

The main elements of the geocoding procedure enable LTR to meet the requirements of Louisiana's legislative mandate (R.S. 40:1105.10) for annual cancer incident reporting by census tract, while also supporting broader research and public health initiatives that rely on high-quality geospatial data. The address normalization and standardization process serve as the foundation on which the remaining three elements stand. Pre-process address standardization takes place before geocoding and focuses on the points in the workflow where address information is captured. This process begins at the facility and includes reviewing information coming from electronic medical records (EMRs) and APIs that will be captured in hospital registry software. This incoming data should be compared, and the most accurate information should be selected by the central registry staff during editing and case consolidation. When reviewing addresses during consolidation, it is essential to remember the anatomy of a US address and focus on the points at which errors are most likely to be introduced. These include discrepancies in numbers, abbreviations, highways, and directional issues, as well as supplemental information, such as names of institutions. Registry staff should incorporate local knowledge in identifying addresses or naming conventions unique to their catchment area.

The post-standardization process takes place when an address has been sent through a geocoder, but the information returned lacks clarity. This includes when the geocoder is unable to identify a location and returns a NAACCR Census Tract Certainty Code (#369) of 4 or higher, which includes census tracts based on zip code and PO box, or when information is missing. LTR incorporates several tools for correcting an address that is unrecognized by the geocoder, which includes searching the individual in a fee-for-service people search software provider, such as LexisNexis-Accurint or Melissa.com. Address standardization is often as simple as paying attention to the spelling of the street names and verifying that the correct zip code is assigned to the city listed in the address. Resources such as the United States Postal Service website, USPS.com, can assist with zip code checks.

Analyzing geospatial data derived from a PO box is never recommended, as it lacks the precision required for meaningful spatial analysis. Cancer registries should exhaust all available options to obtain a physical street address whenever possible. Although not always achievable, an in-depth cleaning of PO box information is recommended. LTR tries to intercept these data during both the pre- and post-process stages of address standardization. To encourage data quality improvements at the source, LTR provides an annual data quality indicator report to hospital registries. Facilities are evaluated based on various data indicators and awarded gold, silver, and bronze certificates. One of the indicators is PO box address at diagnosis, with a benchmark of <5%. By including this geocoding quality indicator, LTR aims to encourage facilities to submit more complete and accurate physical street addresses during the initial data abstraction and pre-processing phase.

Although capturing an address during pre-processing is ideal, it is not always feasible. To address cases in which only a PO box is reported or street-level information is otherwise unavailable, LTR has established a set of guidelines to locate a residential address. These guidelines require a registry to have access to a fee for service people-search database (Accurint) as well as government records such as voter registration and/or driving records. LTR assigns a street level address from Accurint if the PO box address from within the registry software is in Accurint for the correct date range (within 12 months) and a street level address is also present within the same date range (within 12 months). A determination is made if the street level address listed is located directly in the same town or parish/county; or the street level address is in a town or parish/county known to be directly adjacent to the PO box. If voter registration records are available to a registry and the individual is registered to vote and voted in an election in the year of diagnosis, LTR assigns the residential address from voting records if the mailing address matches the PO box reported in the registry records.

Because delayed reporting can occur in cancer registration, LTR focuses on locating an address at diagnosis when it was unavailable at the time of original abstraction. To assign the address at diagnosis for older cases, it is important to start with any associated NAACCR abstract from around the date of diagnosis. Other consolidated records from the same date or the same diagnosis and documentation from the diagnosing facility can also be helpful. The steps previously outlined for obtaining an address for a PO box will also be utilized.

Although not every jurisdiction will agree, LTR has determined that it is appropriate to accept an address from the state death file, with a few caveats. There are usually three addresses of interest listed on a Louisiana death

certificate. These include the residence of the decedent, the informant's address, and the place of death, which may be a separate medical facility. LTR considers it appropriate to assign the decedent's residence as the address at diagnosis if the date of death falls within one year of the cancer diagnosis date. However, extra caution should be taken to avoid assigning the address of a medical facility or family member/friend, as the informant's residence may not be where the decedent was living at the time of their diagnosis.

Conclusion

This report outlines the Louisiana Tumor Registry's procedures for improving address quality in cancer registration, with a focus on geocoding challenges such as PO box reporting, delayed abstraction, and incomplete street-level data. While no outcome data are presented, the described workflows provide a practical framework for registries seeking to enhance geospatial accuracy. These methods, centered on pre- and post-processing, external data sources, and local knowledge, can support more reliable census tract-level mapping and inform future evaluation efforts.

References

1. Mobley LR, Kuo TM. Geographic and Demographic Disparities in Late-stage Breast and Colorectal Cancer Diagnoses Across the US. *AIMS Public Health*. 2015;2(3):583–600. doi: 10.3934/publichealth.2015.3.583. Epub 2015 Aug 28.
2. Nicholson LM, Leider J, Chriqui JF. Exploring the linkage between activity-friendly zoning, inactivity, and cancer incidence in the United States. *Cancer Epidemiol Biomarkers Prev.* 2017 Apr;26(4):578–586. doi:10.1158/1055-9965.EPI-16-0331. Epub 2017 Mar 7.
3. Ghetian CB, Parrott R, Volkman JE, Lengerich EJ. Cancer registry policies in the United States and geographic information systems applications in comprehensive cancer control. *Health Policy.* 2008 Aug;87(2):185–93. doi:10.1016/j.healthpol.2007.12.007. Epub 2008 Feb 19.
4. Parrott R, Volkman JE, Lengerich E, Ghetian CB, Chadwick AE, Hopfer S. Using geographic information systems to promote community involvement in comprehensive cancer control. *Health Commun*. 2010 Apr;25(3):276–85. doi:10.1080/10410231003711755.
5. Shusted CS, Barta JA, Nguyen A, Wen KY, Juon HS, Zeigler-Johnson C. Characterizing lung cancer burden among Asian-American communities in Philadelphia. *J Racial Ethn Health Disparities*. 2024 Oct;11(5):2583–2595. doi:10.1007/s40615-023-01723-1. Epub 2023 Aug 4.
6. Katner, A (2022). St. John the Baptist Parish Chloroprene Monitoring Demonstration Subproject of "Cancer Reporting in St. John the Baptist Parish (CRISP)" Project. Louisiana Department of Health Office of Public Health. **https://louisianacancer.org/st-john/**
7. Wu XC, Maniscalco L, Zhang L, Yi Y, Lefante C, Ricks L, Straif-Bourgeois S, Hsieh MC (eds). *Cancer Incidence in Louisiana by Census Tract, 2006–2014*. New Orleans: Louisiana Tumor Registry, 2018.

How I Did It

Integrating Cancer Registry and Medicaid Data for Survivorship-Focused Research

Siran M. Koroukian, PhD[a,b]; Jamie Shoag, MD, MPH[c]; Long Vu, MS[a]; Johnie Rose, MD, PhD[d]

Abstract: The Ohio Cancer Incidence Surveillance System (OCISS) is not part of the Surveillance, Epidemiology, and End Results (SEER) program; therefore, we are unable to leverage the linked data products using the SEER as the backbone to study survivorship in the state of Ohio. Accordingly, we developed a survivorship-focused data infrastructure integrating 1996–2020 data from the OCISS and 2016–2020 Medicaid data, allowing us to examine health care patterns and outcomes in cancer survivors with short-, medium-, and long-term follow-up, defined as those diagnosed with cancer in the years 2016–2020, 2006–2015, and 1996–2005, respectively. Our integrated database included 95,726 cancer survivors and 2,143,554 cancer-free individuals enrolled in Ohio Medicaid during the 2016–2020 period. Only 9.3% of the study population were diagnosed at the age of 29 or younger, while 32.2% and 58.6% were 30–49 and 50–65 years of age, respectively, when they were diagnosed. Of those diagnosed at age 18 years or younger (n=3,501), 33.0% had long-term follow-up, 43.4% medium-term follow-up, and 23.5% had short-term follow-up. Women with history of breast cancer represented 25.9% and 20.7% of long- and medium-term follow-up, respectively, while individuals with gastrointestinal cancers represented 18.0% of short-term follow-up. Survivors with long-term follow-up included 969 men with prostate cancer. The OCISS-Medicaid database provides unique opportunities to address a broad range of questions in the realm of cancer survivorship and health services research and to advance our understanding of the quality of cancer survivorship care among individuals insured by Medicaid, including those with physical and/or mental disabilities with complex socioeconomic and healthcare needs.

Key words: *Medicaid, Ohio, Ohio Cancer Incidence Surveillance System (OCISS), Surveillance, Epidemiology, and End Results (SEER)*

Introduction

The Ohio Cancer Incidence Surveillance System (OCISS), Ohio's central cancer registry, is not part of the Surveillance, Epidemiology, and End Results (SEER) program. Therefore, the rich linked data products that use the SEER as their backbone cannot be leveraged for cancer outcome studies that are specific to Ohio.

To assess cancer burden in Ohio, the research community in the State of Ohio has been integrating state cancer registry and administrative data. With ongoing improvements in survival and the growing need to examine survivorship in the long term, we sought to develop an integrated database that would provide insight into survivorship outcomes with short-, medium-, and long-term follow-up.

This study focuses on cancer survivors insured by Medicaid—a program considered a "lifeline" by cancer survivors.[1] Indeed, Medicaid represents 1 in 3 children and 1 in 10 adults with a history of cancer.[1] With an average monthly caseload of 3 million individuals, the Ohio Medicaid program provides healthcare coverage to children and pregnant women; low-income parents; aged, blind and disabled persons; and adults under age 65 with incomes up to 138 percent of the FPL.[2] Given that individuals who are older adults (age > 65) represent less than 10% of individuals insured by Medicaid,[3] and that Medicaid claims data alone may not be a complete source of data for the dually eligible, we limited our study population to individuals younger than age 65.

Several factors motivated us to examine health care patterns and outcomes in this population. First, despite their complex socioeconomic and healthcare needs, there is a dearth of survivorship studies on individuals insured by Medicaid. Second, because of financial toxicity, or financial burden caused by medical costs,[4] cancer survivors are more likely than their cancer-free counterparts to find themselves in low-income brackets, and are therefore over-represented among Medicaid beneficiaries. Lastly, the Medicaid program disproportionately serves individuals with chronic conditions (e.g., HIV and developmental disabilities). Taken

---

Drs. Koroukian and Shoag are co-first authors.

[a]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH; [b]Population Cancer Analytics Shared Resource, Case Comprehensive Cancer Center, Cleveland, OH; [c]Pediatric Hematology Oncology and Bone Marrow Transplant, Cleveland Clinic, Cleveland, OH; [d]Center for Community Health Integration, Case Western Reserve University School of Medicine, Cleveland, OH; Population Cancer Analytics Shared Resource, Case Comprehensive Cancer Center, Cleveland, OH

Address correspondence to: Siran M. Koroukian, Ph.D., Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-4945, USA

Cancer incidence data used in these analyses were obtained from the Ohio Cancer Incidence Surveillance System (OCISS), Ohio Department of Health (ODH), a cancer registry partially supported by the National Program of Cancer Registries at the Centers for Disease Control and Prevention (CDC) through Cooperative Agreement Number NU58DP007097. Use of these data does not imply that ODH or CDC agrees or disagrees with the analyses, interpretations, or conclusions in this publication.

This study was funded by Cleveland Clinic's Velosano grant to Dr. Shoag.

Table 1. Characteristics of the Study Population				
Variables of Interest	*Overall*	*Years of Cancer Diagnosis*		
		1996–2005	**2006–2015**	**2016–2020**
Total, *n*	95 726	15 386	39 121	41 219
Median age at cancer diagnosis [IQR]	52.00 [43.00, 59.00]	46.00 [37.00, 55.00]	51.00 [42.00, 57.00]	55.00 [46.00, 60.00]
Age group at cancer diagnosis, *n* (%)				
0–18	3501 (3.7)	1155 (7.5)	1522 (3.9)	824 (2.0)
19–29	5347 (5.6)	1214 (7.9)	2352 (6.0)	1781 (4.3)
30–49	30 793 (32.2)	6761 (43.9)	13 240 (33.8)	10 792 (26.2)
50–65	56 085 (58.6)	6256 (40.7)	22 007 (56.3)	27 822 (67.5)
Median age during years 2016–2020 [IQR]	58.00 [49.00, 64.00]	64.00 [55.00, 73.00]	59.00 [50.00, 65.00]	56.00 [48.00, 61.00]
Age During Years 2016–2020, *n* (%)				
0–18	1502 (1.6)	113 (0.7)	722 (1.8)	667 (1.6)
19–29	3498 (3.7)	625 (4.1)	1304 (3.3)	1569 (3.8)
30–49	19 038 (19.9)	1917 (12.5)	7466 (19.1)	9655 (23.4)
50–65	55 331 (57.8)	5747 (37.4)	21 664 (55.4)	27 920 (67.7)
Over 65	16 357 (17.1)	6984 (45.4)	7965 (20.4)	1408 (3.4)
Length of follow-up, in years (median [IQR])	6.00 [2.00, 11.00]	19.00 [17.00, 21.00]	8.00 [6.00, 11.00]	2.00 [1.00, 4.00]
Length of enrollment in Medicaid, years (mean (SD))	3.11 (1.74)	3.16 (1.75)	3.18 (1.77)	3.02 (1.70)
Race, *n* (%)				
White	72 425 (75.7)	12 084 (78.5)	29 168 (74.6)	31 173 (75.6)
Black	20 965 (21.9)	2995 (19.5)	8987 (23.0)	8983 (21.8)
Other	2336 (2.4)	307 (2.0)	966 (2.5)	1063 (2.6)
Hispanic ethnicity, *n* (%)	1598 (1.7)	177 (1.2)	611 (1.6)	810 (2.0)
Sex = male, n (%)	38 426 (40.1)	4823 (31.3)	14 890 (38.1)	18 713 (45.4)
SEER summary stage, *n* (%)				
In situ	6341 (6.6)	1644 (10.7)	2824 (7.2)	1873 (4.5)
Localized	42 673 (44.6)	8290 (53.9)	18 652 (47.7)	15 731 (38.2)
Regional	23 549 (24.6)	2982 (19.4)	9671 (24.7)	10 896 (26.4)
Distant	18 139 (18.9)	1309 (8.5)	5876 (15.0)	10 954 (26.6)
Unstaged	3383 (3.5)	1124 (7.3)	985 (2.5)	1274 (3.1)
Missing	1641 (1.7)	37 (0.2)	1113 (2.8)	491 (1.2)
County of residence, *n* (%)				
Metro	65 168 (68.1)	10 230 (66.5)	26 977 (69.0)	27 961 (67.8)
Appalachia	19 403 (20.3)	3328 (21.6)	7712 (19.7)	8363 (20.3)
Non-Metro	11 155 (11.7)	1828 (11.9)	4432 (11.3)	4895 (11.9)
Marital status, *n* (%)				
Married	18 983 (19.8)	3067 (19.9)	7719 (19.7)	8197 (19.9)
Missing	13 194 (13.8)	1993 (13.0)	5023 (12.8)	6178 (15.0)
Single	63 549 (66.4)	10 326 (67.1)	26 379 (67.4)	26 844 (65.1)

together, these factors provided a strong rationale for our effort to create a survivorship-focused data infrastructure specific to Ohio's most vulnerable. While there will be many uses of this database in cancer- and survivorship-related studies pertaining to socioeconomically vulnerable populations, this article presents an initial assessment of the population represented herein by demographics, cancer type, and length of follow-up.

Methods

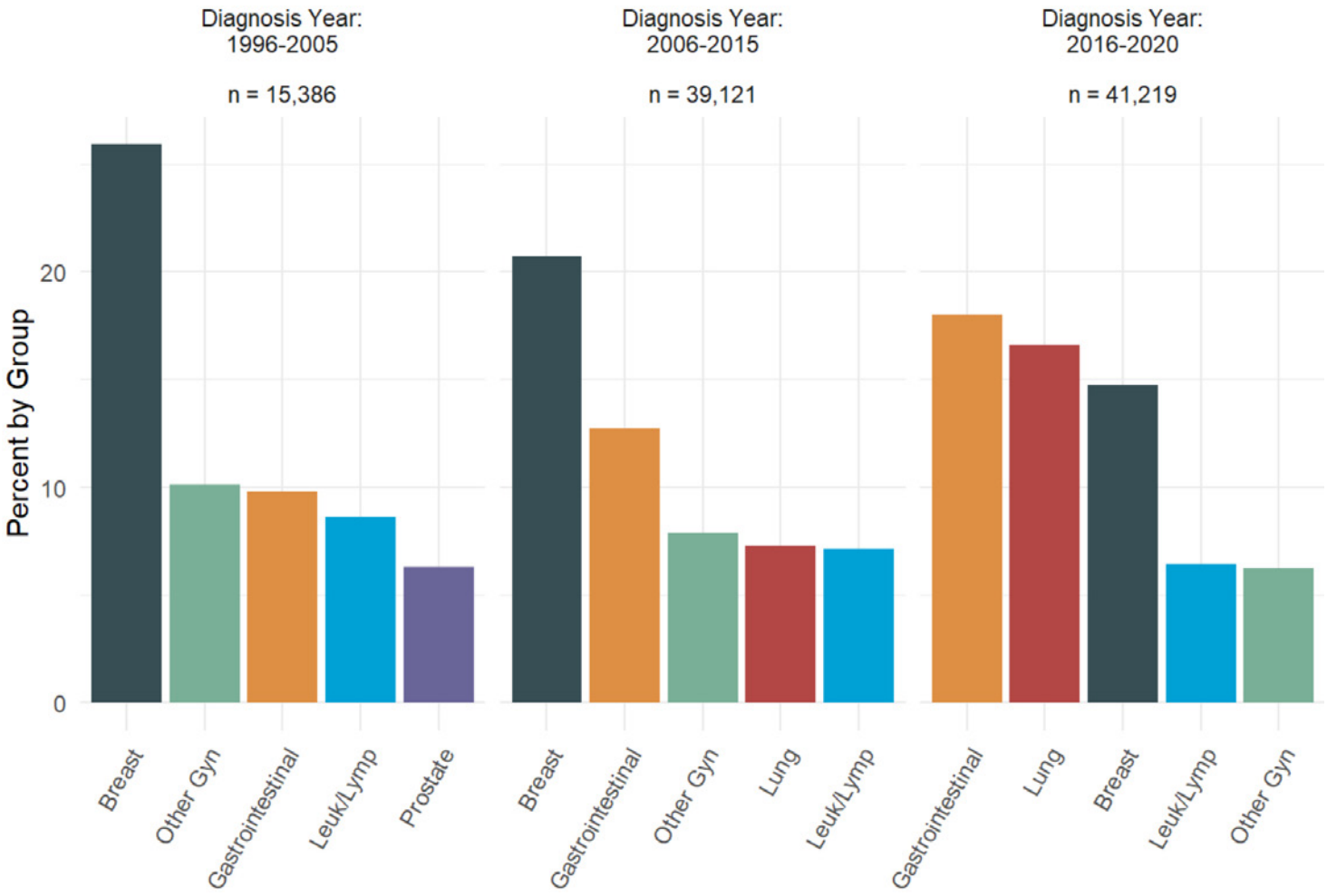
Our data resource consisted of integrated OCISS and Medicaid data. The OCISS data included all patients diagnosed with cancer during the years 1996–2020, regardless of their cancer type, while the Medicaid enrollment and claims files covered the period 2016–2020. The limited span of Medicaid data was dictated by budgetary constraints. Using a deterministic linkage algorithm, the data were linked by the Centers for Medicare & Medicaid Services (CMS), using Social Security number (SSN), date of birth, and sex—the three data elements required by CMS in SSN finder files. These data were sent to CMS by the Ohio Department of Health. The files obtained from CMS included encrypted OCISS patient and Medicaid IDs.

The OCISS includes data on newly diagnosed cancer cases among Ohio residents. As with other central registries, incident cases of in situ cervical cancer, as well as squamous- and basal-cell carcinoma of the skin, are not required to be reported to the OCISS. In recent years, the completeness of the OCISS in a given year has been estimated to be in the high 90%.

In addition to patient demographics, health insurance, and residence address at cancer diagnosis, OCISS variables include anatomic cancer site, cancer stage, date at treatment initiation, treatment modalities, as well as vital status and date and cause of death, when applicable.

Medicaid enrollment files include a single record for each individual enrolled in the Medicaid program in a given calendar year. Enrollment records include demographic variables and monthly enrollment status and eligibility data. For individuals diagnosed with cancer during the years 2016–2020, monthly enrollment data allow us to construct enrollment history in Medicaid relative to their cancer diagnosis, as previous studies have shown the importance of the timing of enrollment in Medicaid relative to cancer diagnosis. In fact, compared to cancer patients enrolled in Medicaid several months prior to cancer diagnosis, those

Figure 1. Top 5 Cancer Types Among Cancer Survivors, by Year of Cancer Diagnosis



enrolling in Medicaid at around the time of cancer diagnosis are more likely to be diagnosed with more advanced disease, especially in the case of screening-amenable cancers.[5–7] For individuals diagnosed with cancer before 2016, enrollment data allow us to ascertain individuals' Medicaid status during the years 2016–2020.

Medicaid claims files include inpatient claims data, outpatient institutional and non-institutional claims data, and pharmacy data. Inpatient and outpatient data carry the date of service, as well as diagnosis and procedure codes, allowing us to identify comorbid conditions, services received while using pharmacy data, we can identify use of oral medications and polypharmacy.

Lastly, we note that our database was augmented considerably by including data for randomly selected cancer-free Ohio residents insured by Medicaid during the years 2016–2020. These individuals will serve as controls in our studies. Again, because of budgetary considerations, we limited the number of controls to 3 times the number of total (Medicaid and non-Medicaid) cancer cases (n= 2,143,554).

This study was approved by the Case Western Reserve University (#STUDY20221103), the Cleveland Clinic IRB (#22-1054), the Ohio Department of Health (#2023-02), and the Centers for Medicare & Medicaid Services (RSCH-2023-59092).

Results

Overall, the database included 95,726 individuals diagnosed with incident cancer during the years 1996–2020 and enrolled in Medicaid during the years 2016–2020 (Table 1). Of those, 16.1% were diagnosed in the years 1996–2005 (hereafter referred to as "survivors with long-term follow-up"), 40.9% were diagnosed in the years 2006–2015 ("medium-term follow-up"), and 43.1% were diagnosed in the years 2016–2020 ("short-term follow-up"). The median age at diagnosis [and interquartile range (IQR)] in these three categories were 46 [37,55], 51 [42,57], and 55 [46,60] years, respectively. The median length of follow-up was 19 [17,21], 8 [6,11], and 2 [1,4] years, respectively, in these three groups. The percentage of males was highest among survivors with short-term follow-up (45.4%), and lowest among those with long-term follow-up (31.3%). Individuals identified as White represented 75.6% of survivors with short-term follow-up and 78.5% of those with long-term follow-up, while the percentage of Black patients ranged between 19.5% and 23.0% among survivors with long- and medium-term follow-up, respectively. The percent of individuals categorized as having Hispanic ethnicity ranged from 1.2% to 2.0%.

Only 9.3% of the study population were diagnosed at the age of 29 or younger, while 32.2% and 58.6% were 30–49 and 50–65 years old at diagnosis, respectively. Of those diagnosed with cancer at age 18 or younger (n=3,501), 33.0%, 43.4%, and 23.5% with long-, medium-, and short-term follow-up, respectively.

Figure 1 shows the top 5 types of cancer within each of the long-, medium-, and short-term follow-up survivor groups. Women diagnosed with breast cancer represented 25.9% and 20.7% of survivors with long- and medium-term follow-up, respectively. Patients with leukemia represented 6.4% to 8.6% of survivors with long- and medium-term follow-up, respectively. The percentage of patients with gastrointestinal cancers was highest among survivors with short-term follow-up (18.0%), and lowest among those with long-term follow-up (9.8%). Women with gynecological cancers represented 10.1% of survivors with long-term follow-up, and 6.3% of those with short-term follow-up. Lastly, we note that survivors with long-term follow-up included 969 men with a history of prostate cancer.

Discussion

This database presents unprecedented opportunities to gain a better understanding of the experience of survivors insured by Medicaid with short-, medium-, and long-term follow-up among subgroups defined by their co- or multimorbidity profile, use of cancer and non-cancer screening services, and healthcare utilization—to mention just a few areas of investigation. Importantly, given that 1 in 3 childhood cancer survivors are insured by Medicaid, we are also able to study downstream outcomes in children who have experienced cancer and its treatment, including developmental delays.

An additional unique opportunity presented by this database is the study of cancer survivorship in special populations, including (but not limited to) those with HIV, severe mental illness, and various physical and/or intellectual disabilities—subgroups of the population that remain understudied but are over-represented in the Medicaid-insured population.

This database also comes with limitations, in addition to lacking patient-centered data. First, this database is limited to the state of Ohio and to those insured by Medicaid; therefore, the results may not be generalizable to cancer survivors not insured by Medicaid, or those insured by Medicaid but residing in other states. Second, as noted earlier, due to budgetary constraints, we were unable to purchase Medicaid data for more than 5 years. This prevented us from being able to examine cancer patients' history of enrollment in Medicaid before and after cancer diagnosis for people diagnosed with cancer before 2016. Third, the OCISS includes both insurance status and residence address at the time of diagnosis, while the Medicaid data ascertain Medicaid status and provide the census tract of residence in the 2016–2020 period. For individuals diagnosed in the earlier years of the study period, we do not have data on temporal changes in their insurance coverage, residential history, or social determinants of health. Similarly, it is not possible to assess their use of health services except during 2016–2020.

Despite these limitations, the OCISS-Medicaid database provides unique opportunities to address a broad range of questions on cancer survivorship and to advance our understanding of the quality of cancer survivorship care among individuals insured by Medicaid. In the absence of claims data in the SEER-Medicaid database, databases such as this one address an important knowledge gap in the survivorship literature.

References

1. Canceradvocacygroup.org. Protecting Access to Medicaid for Cancer Survivors. **https://canceradvocacy.org/policy/protecting-access-medicaid-cancer-survivors/**, Accessed September 29, 2025.
2. Ohio Department of Medicaid, Caseload Summary. **https://tinyurl.com/yym6fca8**, Accessed January 3, 2026.
3. Kaiser Family Foundation, Medicaid Enrollees by Age. **https://tinyurl.com/4hc78ftm**, Accessed January 3, 2026.
4. Lee KL, Eniu A, Booth CM, MacDonald M, Chino F. Financial toxicity and breast cancer: Why does it matter, who is at risk, and how do we intervene? *Am Soc Clin Oncol Educ Book*. Jun 2025;45(3):e473450. doi:10.1200/EDBK-25-473450
5. Koroukian SM, Dong W, Albert JM, et al. Post-Affordable Care Act improvements in cancer stage among Ohio Medicaid beneficiaries resulted from an increase in stable coverage. *Med Care*. Nov 1 2022;60(11):821–830. doi:10.1097/MLR.0000000000001779
6. Bradley CJ, Given CW, Roberts C. Late stage cancers in a Medicaid-insured population. *Med Care*. Jun 2003;41(6):722–8. doi:10.1097/01.MLR.0000065126.73750.D1
7. Koroukian SM. Assessing the effectiveness of Medicaid in breast and cervical cancer prevention. *J Public Health Manag Pract*. Jul-Aug 2003;9(4):306–14.

Short Report

Investigation of High Number of Primary Tumors Reported in Central Sequence Number

Rebecca Moore[a]; Jeanie Hartley[b]; Fran Maguire[c]; Recinda Sherman[d]

Background

A Central Sequence Number records the sequence of reportable neoplasms during the patient's lifetime as determined (**Sequence Number-Central**, North American Association of Central Cancer Registries [NAACR] Item #380) by the central registry. Valid values range from "00" for one primary in a patient's lifetime, to "59" for the 59th primary and higher, while the values of "60" to "87" are reserved for nonmalignant cases. There was some concern that certain cancer sites, such as in situ cutaneous melanoma, could have high numbers of primary tumors going beyond 59. For such situations, the coding rules state to assign sequence number "59" to all primaries of "59" and above. However, not all central registry software systems allow multiple entries of "59" for a given patient.

Objective

A prominent central registry software vendor was concerned about the programmatic problem of having more than 59 malignant tumors. They reached out to NAACCR to investigate the scope of the potential problem in the Cancer in North America (CiNA) data (**https://www.naaccr.org/cina-data-products-overview/**). Members from a NAACCR Research and Data Use work group, the CiNA Writing Network, investigated sequence numbers of 55 or more in CiNA data and conducted follow-back with registries to determine if the high sequence number was in error and to assess their knowledge about coding sequences of >59.

Methods

We queried the CiNA dataset in SEER*Stat for cases diagnosed between 1995 and 2022 with central sequence numbers of 55 or higher. We identified the registry, tumor site, year of diagnosis, and histologic type. We reached out to each identified registry through email to ask whether the sequence number was correct and what the registry uses as the protocol for coding cases with sequence numbers of > 59.

Results

We identified 10 cases with sequence numbers of 55 or greater from five registries. Six were melanomas, 1 was urinary bladder, 1 was prostate, and 2 were breast. Two registries were using the correct number for sequence codes, but 3 registries were not. The staff at 2 registries were aware that multiple primaries above sequence 59 should all be coded as 59. The other 3 registries stated that they had coded primaries with greater than 59 sequences but were not aware of the coding rules that address this situation. One registry conducted an evaluation on its full dataset and found 8 additional cases with miscoded sequence numbers for 55 or greater.

Discussion

Due to this evaluation, an edit has been made to allow multiple values of sequence 59 per patient, which ensures that all registry software systems will also allow multiple values of 59. Because the majority of the "55" or greater sequence numbers were in error, and because an additional 8 cases were found by the central registry that reviewed its full dataset, we recommend that every registry run a query for high sequence numbers for review.

Identification of this issue has also prompted wider discussion in the NAACCR community. Based on these discussions, the Solid Tumor Manual Editorial Board will review the multiple primary rules for melanoma to evaluate whether any changes need to be made.

This project underscores the benefits of integration between standards and operations and data use in our field. A set of analysts defined the problem in a research dataset, and a set of abstractors conducted follow-back to understand the underlying issues. Cancer surveillance functions optimally when analysis and data collection activities are integrated.

[a]Cincinnati Children's Hospital, 3333 Burnet Avenue, Cincinnati, OH; [b]ODS Medovent Solutions, St. Petersburg, FL; [c]UC Davis Health/California Cancer Registry, Sacramento, CA; [d]North American Association of Central Cancer Registries, Claremont, CA

Address correspondence to Rebecca Moore, Cincinnati Children's Hospital, Cincinnati, OH, Rebecca.Moore2@cchmc.org.

Association of Distance to Diagnosis and Area-Based Social Measures with Stage at Diagnosis among Iowans with HPV-Related Cancers

Emily Janio, MPH[a]; Amanda Kahl, MPH[b]; Natoshia Askelson, MPH, PhD[a]; Sarah Nash, MPH, PhD[b]

Introduction

Human papillomavirus (HPV) causes oral and anogenital cancers, the incidence of which is increasing.[1] Late-stage diagnosis is associated with lower survival.[2] Timely diagnosis may be impeded by lower neighborhood-level socioeconomic status (SES)[3,4] and greater distance to place of diagnosis.[5]

Iowa is an ideal location to understand the association between distance, SES, and stage at diagnosis among patients with HPV-associated cancers, due to its rurality[6] and increasing incidence of HPV-related cancers.[7]

Objectives

1. To assess whether distance is associated with stage at diagnosis irrespective of area-based social measure (ABSM) controlled for.

2. To assess whether select ABSMs are associated with stage at diagnosis.

3. To assess whether each ABSM independently moderates the association between distance and stage at diagnosis.

Methods

We used data on adults with HPV-associated cancers (cervical, oropharyngeal, vulvar/vaginal, and anal/rectal) from the Iowa Cancer Registry from 2010–2021.

Neighborhood-level SES was operationalized using 4 census tract-level ABSMs.

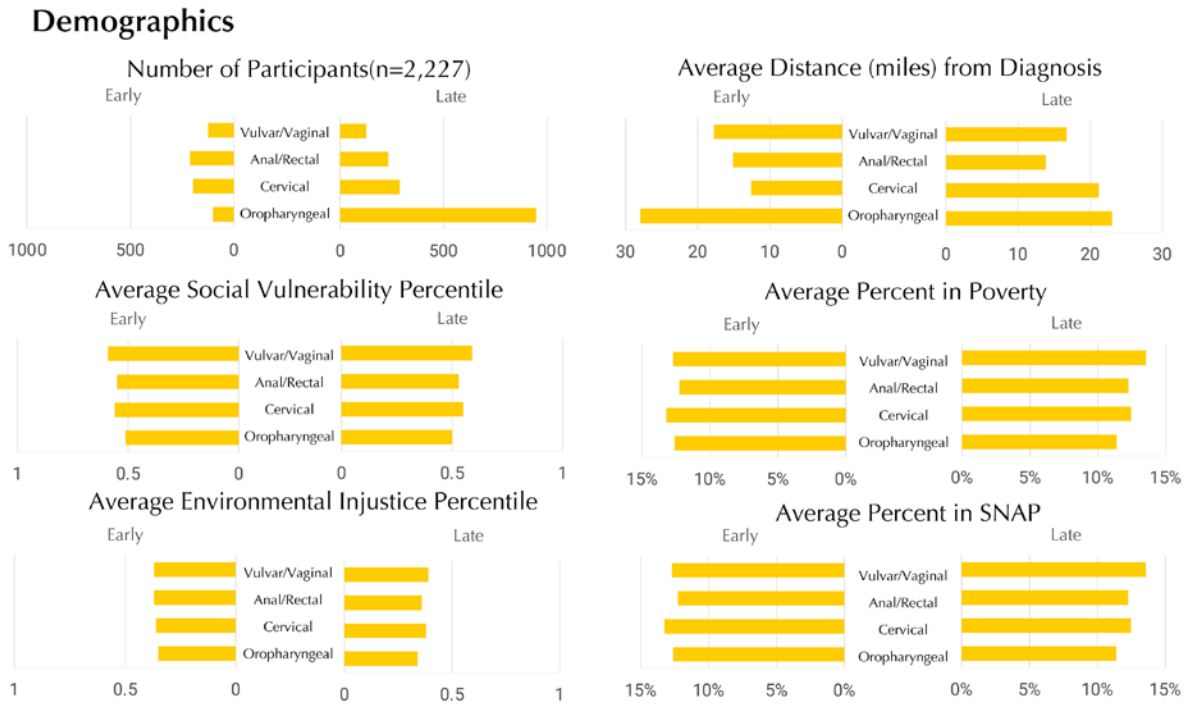
Distance was measured from the centroid of a person's census tract to their reporting facility using ArcGIS (Figure 1).

For each cancer site, we ran 4 logistic regression models predicting stage at diagnosis and controlling for distance, one of the 4 ABSMs, and sociodemographic factors (Figure 2).

Conclusions

While the association between distance and stage among those with cervical cancer was significant, the relative weakness of this association suggests that it may not be highly clinically relevant.

Figure 1. Data from the Iowa Cancer Registry from 2010–2021

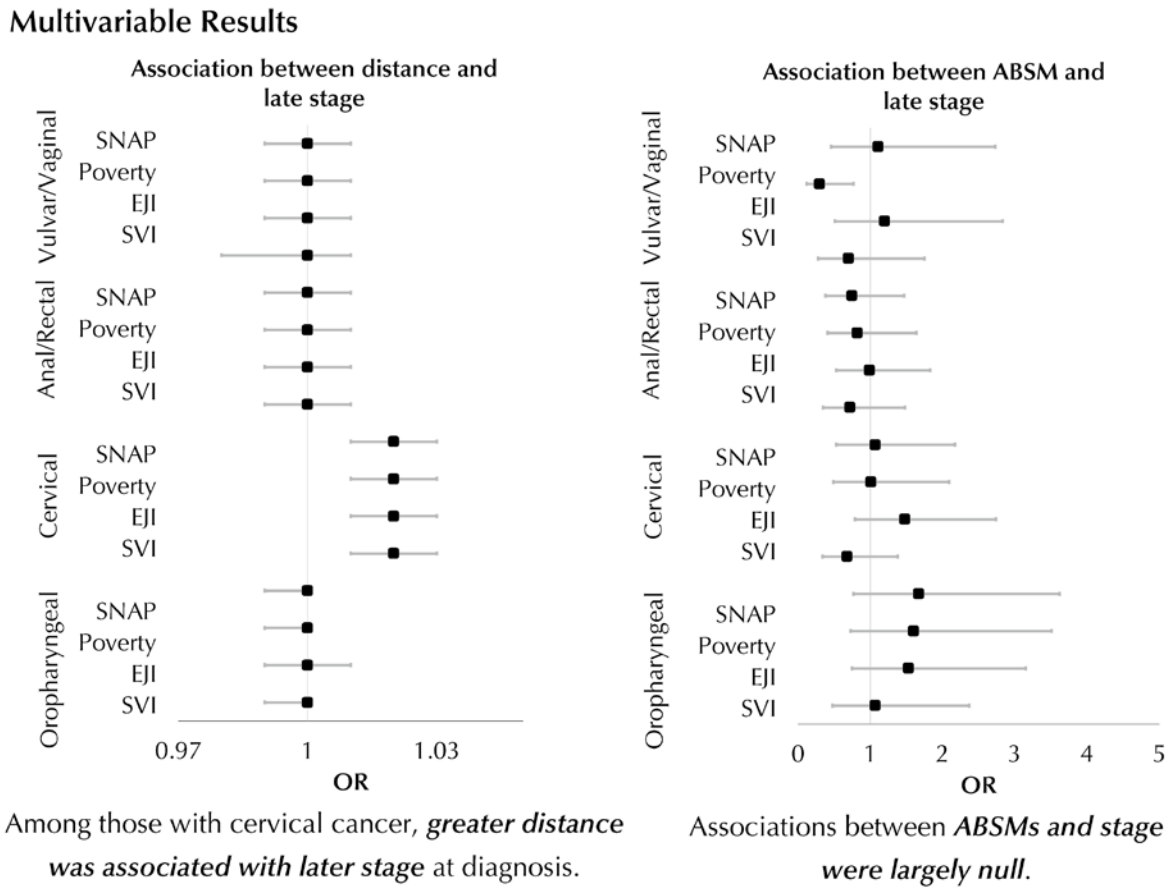


Registry contact: Iowa Cancer Registry (**https://shri.public-health.uiowa.edu/**); Iowa Cancer Dashboard (**https://shri.public-health.uiowa.edu/cancer-data/**)

[a]University of Iowa College of Public Health, Department of Community and Behavioral Health, bUniversity of Iowa College of Public Health, Department of Epidemiology

This content was originally presented as a poster at NAACCR's Annual Conference in Hartford, Connecticut in June 2025.

Figure 2. Results of Logistical Regression Models



The differences in the associations between distance and stage according to cancer type mirrors previous studies among Iowans.[5,8]

The largely null associations could support a "true" lack of association or could reflect limitations of census-level ABSMs, which do not reflect individual-level household SES.

This exploratory study suggests the importance of considering area-level SES when identifying those most at risk of untimely care.

Acknowledgements

Dr. Nash and Ms. Kahl are supported by the Iowa Cancer Registry (Contract HHSN261201300101, Task Order HHSN26100005) supported by the NCI's Surveillance, Epidemiology and End Results Program. Drs. Nash and Askelson are supported by the University of Iowa Holden Comprehensive Cancer Center 3P30CA086862. Ms. Janio is supported by the University of Iowa Medical Scientist Training Program NIH grant T32 GM139776 and Gift Funds of The Holden Comprehensive Cancer Center at the University of Iowa.

References

1. Van Dyne EA, Henley SJ, Saraiya M, Thomas CC, Markowitz LE, Benard VB. Trends in Human Papillomavirus–Associated Cancers — United States, 1999–2015. *MMWR Morb Mortal Wkly Rep*. 2018;67(33):918–924.
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12–49.
3. Boscoe FP, Henry KA, Sherman RL, Johnson CJ. The relationship between cancer incidence, stage and poverty in the United States. *Int J Cancer*. 2016;139(3):607–612.
4. Boscoe FP, Johnson CJ, Sherman RL, Stinchcomb DG, Lin G, Henry KA. The relationship between area poverty rate and site-specific cancer incidence in the United States. *Cancer*. 2014;120(14):2191–2198.
5. Henry KA, Sherman R, Farber S, Cockburn M, Goldberg DW, Stroup AM. The joint effects of census tract poverty and geographic access on late-stage breast cancer diagnosis in 10 US States. *Health Place*. 2013;21:110–121.
6. *Report to the Governor of Iowa and the General Assembly on "Physician Workforce Study" As Requested in House File 532, Section 4*. University of Iowa Carver College of Medicine Office of Statewide Clinical Education Programs; 2020.
7. National Cancer Institute. Incidence Rates Table. State Cancer Profiles. Accessed October 31, 2024.
8. Charlton ME, Matthews KA, Gaglioti A, et al. Is travel time to colonoscopy associated with late-stage colorectal cancer among Medicare beneficiaries in Iowa? *J Rural Health*. 2016;32(4):363–373.

2025 NAACR Standards and Registry Operations Winner

New Jersey State Cancer Registry Data Ascertainment for Melanoma Research Studies

Eileen Flores, DrPH, MPH[a]; Katelyn Bleeker[b]; Brian Cahill, BA, CTR[a]; Heather Stabinsky, MSEd, ODS-C[b]; Lisa Paddock, PhD[b,c]; Antoinette Stroup, PhD[b,c]

Background

Data quality varies across reporting facilities. Hospitals are staffed with certified oncology data specialists (ODS-C) trained in cancer data collection, whereas non-hospital reporting facilities generally rely on non-ODS personnel to report cancer data. As a result, reporting from non-hospital facilities is often inconsistent and may be missing key data items such as race, ethnicity and stage at diagnosis. Incomplete data in key fields result in people being unnecessarily excluded from melanoma research studies who may otherwise be eligible.

Objective

The study objective was to quantify and characterize melanoma patients who are excluded for research study recruitment due to unknown race, ethnicity, and stage at diagnosis at the New Jersey State Cancer Registry (NJSCR).

Methods

Data Source

All data came from the NJSCR.

Analytic Cohort

NJSCR was accessed using SEER*DMS to create a cohort of melanoma patient records/sets with commonly used research study selection criteria coded to NAACCR standards: SEER Site Recode = "25010", SEER Reportable = "1", diagnosed 2019–2023 (#390), NJ resident at diagnosis (#1820).

Statistical Methods

Descriptive statistics were used to summarize missing race (#160), ethnicity (#190), and Summary Stage 2018 (SS 2018) (#764) by type of reporting source (#500) among patient records and sets. Missing data was defined as anything coded "99" (race), "9" (ethnicity, SS 2018), or left blank (Figure 1).

Pearson's chi-square tests were used to compare patient records and sets. Statistical significance was set at the 0.05 level.

Data management and analyses were conducted with Microsoft Excel and SAS version 9.4 (SAS Institute Inc, Cary, NC).

Figure 1. Examples of North American Association of Central Cancer Registries (NAACCR) Code Values and Descriptions and Link to the Full NAACCR Data Standards and Data Dictionary

Race (#160)	
Code	Description
1	White
2	Black or African American
3	American Indian or Alaska Native
4-8,10-17,20-22,25- 28,30-32,96,97	Asian
98	Some other race
99	Unknown by patient

Summary Stage 2018 (#764)	
Code	Description
0	In situ
1	Localized only
2	Regional by direct extension only
3	Regional lymph nodes only
4	Regional by BOTH direct extension AND lymph node involvement
7	Distant site(s)/node(s) involved
8	Benign/borderline
9	Unknown if extension or metastasis (unstaged, unknown, or unspecified) Death certificate only case

Link to full NAACCR Data Standards and Data Dictionary

Registry contact: New Jersey State Cancer Registry (**https://www.nj.gov/health/ces/reporting-entities/njscr/**); New Jersey Cancer Dashboard (**https://www.nj.gov/health/ces/cancer-researchers/cancer-data/**)

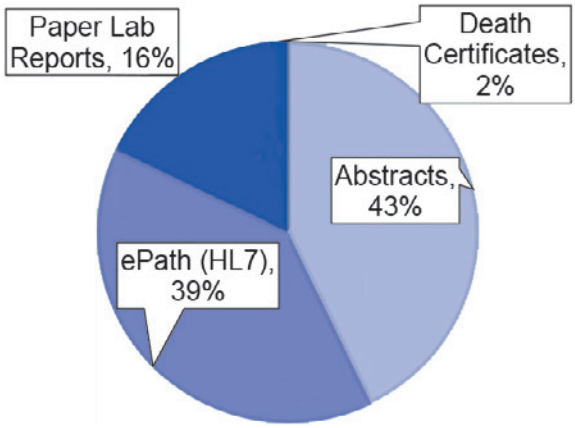
[a]Cancer Epidemiology Services, New Jersey State Cancer Registry, New Jersey Department of Health, Trenton, NJ; [b]Rutgers Cancer Institute, New Brunswick NJ; [c]Rutgers School of Public Health, Piscataway, NJ

This content was originally presented as a poster at NAACCR's Annual Conference in Hartford, Connecticut in June 2025.

Results

A total of 59,824 records meeting the search criteria were identified from SEER*DMS. The majority of record subtypes were abstracts (43%), followed by ePath (HL7, or health level 7) (39%), paper lab reports (16%), and death certificates (2%) (Figure 2).

Figure 2. Summary of Melanoma Records 2019–2023



Among 59,824 records, missing race and ethnicity information was significantly higher in non-hospital records compared to hospital records (62% versus 15%, 95% CI: -0.42- -0.41) (Table 1).

Among 25,639 abstracts, SS 2018 (#764) was missing in 76% (n=6,876) of non-hospital abstracts compared to 7% (n=1,142) of hospital abstracts (95% CI: 0.71-0.74). Among non-hospital abstracts with stage data, 362 abstracts (17%) were completed by an ODS-C. Note that only abstracts were included since SS 2018 is not reported in paper lab reports, ePath (HL7), and death certificates (Table 2).

Melanoma patient sets (n=26,382) are reported with unknown race (7%) more frequently than the thresholds set by NPCR and NAACCR (≤ 3%;≤ 5%) (Figure 3).[1,2]

Conclusions

Ascertaining eligibility criteria for melanoma research study recruitment in a timely manner can pose a challenge depending on the quality of the reporting source. At NJSCR, a significant number of melanoma cases are reported from ePath (HL7), paper lab reports, or Web Plus abstracts from non-hospitals. These records are more likely to be missing race, ethnicity, and summary stage, which must be ascertained from follow-back with physician offices, staged at the central registry, or obtained through linkages. Uncoded data reduces the chance for potentially eligible patients to be included in research, especially Hispanics and non-Hispanic Blacks and Asians,[3] who have lower melanoma-specific survival rates compared to non-Hispanic Whites. NJSCR must explore methods to improve non-hospital reporting for race and ethnicity, as well as extracting staging from records lacking summary stage to maximize study recruitment efforts.

Table 1. Race/Ethnicity by Reporting Source Type (#500)

	HOSPITAL (n=31,175)	NON-HOSPITAL (n=27,702)	DEATH CERTIFICATE (n=948)
RACE/ETHNICITY	n (%)	n (%)	n (%)
+Race / +Ethnicity	16,796 (54)	8,938 (32)	948 (100)
+Race / - Ethnicity	9,204 (30)	1,734 (6)	0 (0)
- Race / +Ethnicity	425 (1)	234 (1)	0 (0)
- Race / - Ethnicity	4,783 (15)	15,892 (62)	0 (0)

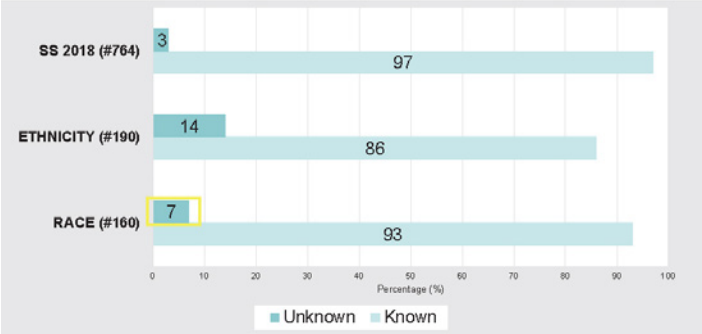
'+' = Known; '-' = Unknown; Statistically Significant

Table 2. Summary Stage by Reporting Source Type (#500)

	HOSPITAL (n=16,630)	NON-HOSPITAL (n=9,009)
	n (%)	n (%)
STAGE	15,488 (93)	2,133 (24)
Certified ODS	15,488	362
Non-Certified ODS	0	1,771
UNSTAGED	1,142 (7)	6,876 (76)
Certified ODS	1,142	7
Non-Certified ODS	0	6,869

Statistically Significant

Figure 3. Summary of Melanoma Cases 2019–2023



Acknowledgements

The New Jersey State Cancer Registry is supported by the National Program of Cancer Registries of the Centers for Disease Control and Prevention under cooperative agreement NU58DP007117-03-00 awarded to the New Jersey Department of Health, the Surveillance, Epidemiology, and End Results program of the National Cancer Institute under contract 75N91021D00009 awarded to the Rutgers Cancer Institute of New Jersey, and the State of New Jersey.

References

1. CDC. National Program of Cancer Registries. 2024. Accessed April 4, 2025. **https://www.cdc.gov/national-program-cancer-registries/about/npcr-standards.html**
2. NAACCR. Certification Criteria. 2018. Accessed April 4, 2025. **https://www.naaccr.org/certification-criteria**
3. Qian, Y., Johannet, P., Sawyers, A., Yu, J., Osman, I., & Zhong, J. (2021). The ongoing racial disparities in melanoma: an analysis of the surveillance, epidemiology, and end results database (1975-2016). *Journal of the American Academy of Dermatology*, *84*(6):1585–1593.

Journal of Registry Management Continuing Education Quiz—WINTER 2025

THE BURDEN OF CANCERS ASSOCIATED WITH HEREDITARY BREAST AND OVARIAN CANCER SYNDROME, 2015–2019: A POPULATION-BASED CANCER REGISTRY ANALYSIS

Recinda Sherman, PhD, MPH, ODS-C

This quiz is derived from the article, "The Burden of Cancers Associated with Hereditary Breast and Ovarian Cancer Syndrome, 2015–2019: A Population-based Cancer Registry Analysis," by Caroline B. Morales, MPH; Lisa E. Paddock, PhD; Katelyn Bleeker, MPH, ODS-C; and Antoinette M. Stroup, PhD.

After reading the article and completing the quiz, participants will be able to:

- Understand the significance of hereditary breast and ovarian cancer syndrome and related cancers.
- Understand the limitations of current estimates of the burden of hereditary breast and ovarian cancers.
- Describe the study-specific burden of hereditary breast and ovarian cancers by stage at diagnosis, race, and ethnicity.

1. Which of the following statements is true of hereditary breast and ovarian cancer syndrome (HBOCS)?
   a) It does not increase the risk of cancer.
   b) It is caused by tobacco use.
   c) It is caused by unknown environmental factors.
   d) It is an inherited condition.
2. Is the burden of hereditary breast and ovarian cancers well understood?
   a) Yes
   b) No
   c) Unknown
3. What percentage of breast cancers are estimated to be attributed to inherited mutations in current literature:
   a) Up to 10%
   b) Up to 15%
   c) Over 35%
   d) None
4. What percentage of ovarian cancers are estimated to be attributed to inherited mutations in current literature?
   a) Up to 10%
   b) Up to 15%
   c) Over 35%
   d) None
5. Is the following statement true, false, or unknown? "Male carriers of breast cancer mutations can pass risk to female descendants but are not at greater risk of breast cancer themselves."
   a) True
   b) False
   c) Unknown
6. How did the results of this study compare to those of prior literature?
   a) It estimated the same proportion of possible HBOCS-related breast cancers as prior literature.
   b) It estimated a lower proportion of possible HBOCS-related breast cancers as prior literature.
   c) It estimated a higher proportion of possible HBOCS-related breast cancers as prior literature.
   d) It was unable to estimate the proportion of possible HBOCS-related breast cancers.
7. Which race/ethnic group had the highest proportion of possible HBOCS-related cancers:
   a) Whites
   b) Non-whites
   c) Blacks
   d) Hispanics
   e) Asian/PIs
   f) It depends upon cancer type.
8. How did incidence by stage at diagnosis for the possible HBOCS-related breast cancers vary from that of breast cancers overall?
   a) It was the same as for as breast cancers overall.
   b) It was higher for local stage than breast cancers overall.
   c) It was higher for distant stage than breast cancers overall.
   d) No differences by stage were statistically significant.
9. How did incidence for possible HBOCS-related cancers vary geographically?
   a) It did not vary by geography in NJ.
   b) It was highest in the northern part of NJ.
   c) It was highest in the southern part of NJ.
   d) It varied by county with no discernable geographic pattern.
10. How was this study limited?
    a) This study did not have limitations on the interpretation of results.
    b) It may underestimate the burden of HBOCS-related cancers.
    c) It may overestimate the burden of HBOCS-related cancers.
    d) This study did not estimate the burden of HBOCS-related cancers.

Purchase Quiz to Earn CE:

1. Go to **http://www.cancerregistryeducation.org/jrm-quizzes**
2. Select quiz and "Add to Cart" (You may be prompted to login using your NCRA login).
3. Continue through the checkout process.
4. Once purchase is complete, the quiz will load automatically into "My Learning Activities" page.

Journal of Registry Management
Volume 52, Spring 2025 to Winter 2025

Reviewer acknowledgement: *JRM gratefully acknowledges the individuals who have served as manuscript reviewers or have otherwise assisted in the review process during the past year. Their wise counsel and contributions to the Journal have been most valued.*

Multiple Author Index—Vol. 52 (2025)

F

Fitts, Austin

Shivanna A, Fitts A, Tschida J, Chandrashekar M, Hwang A, Hsu E, Gounley J, Hanson HA. Leveraging Large Language Models for Real-World Data Evidence: A Framework for Automated Treatment Extraction and Data Harmonization. Winter;52(4):138-145.

Flores, Eileen

Flores E, Bleeker K, Cahill B, Stabinsky H, Paddock L, Stroup A. New Jersey State Cancer Registry Data Ascertainment for Melanoma Research Studies. Winter;52(4):172-173

Francis, Sean

Retell J, Francis S, Rene J, Baade P. CaRDO: A Tool for Empowering Cancer Registries to Publish Population-Level Cancer Statistics Online. Winter;52(4):146-150.

G

Gann, Peter H.

Kato I, Bosland MC, Kim S, Gann PH. Chronological Trends of PSA Levels at Diagnosis in Population-based Prostate Cancer Data 2004-2021, a Time Period of Drastic Changes in Screening Guidelines: Potential Implications in Prostate Cancer Diagnosis. Fall;52(3):70-78

Garza, Maryam Y.

Conner CL, Redding MC, Seker E, Greer ML, Williams TB, Garza MY. Data-Related Processes and Challenges at Level I–IV Trauma Centers in Arkansas: Results of a Mixed-Methods Case Study. Spring;52(1):6-15

Gatta, Praneetha

Bayakly R, McNamara C, Blankenship E, Hunter N, Gatta P. The Georgia Cancer Registry Database Improves Local Cancer Data Access. Fall;52(3):99

Gounley, John

Shivanna A, Fitts A, Tschida J, Chandrashekar M, Hwang A, Hsu E, Gounley J, Hanson HA. Leveraging Large Language Models for Real-World Data Evidence: A Framework for Automated Treatment Extraction and Data Harmonization. Winter;52(4):138-145.

Grannis, Shawn

Wilcox A, Grannis S, Zachary I, Venterisv N, Abrams Z, Newland J, Beheshti M, Ham L, Day J, Barnholtz-Sloan JS. Evaluating Privacy-Preserving Record Linkage in Merging Cancer Registry Data and Electronic Records. Fall;52(3):79-83

Greene, Frederick L.

Greene FL, Dalrymple SF. The Ambiguity of "Ambiguous Terminology." Spring;52(1):4-5

Greer, Melody L.

Conner CL, Redding MC, Seker E, Greer ML, Williams TB, Garza MY. Data-Related Processes and Challenges at Level I–IV Trauma Centers in Arkansas: Results of a Mixed-Methods Case Study. Spring;52(1):6-15

H

Haggan, Kim

Boris K, Neas J, Bancroft C, Haggan K. Veterans' Data Submissions Improve High-Quality Cancer Data in Maine and Address Historic Gaps. Summer;52(2):51

Hailer, Laurie

Chapman SA, Hailer L, Miller J. Workload and Staffing Study of Hospital Cancer Registries. Spring;52(1):16-21

Ham, Lucinda

Wilcox A, Grannis S, Zachary I, Venterisv N, Abrams Z, Newland J, Beheshti M, Ham L, Day J, Barnholtz-Sloan JS. Evaluating Privacy-Preserving Record Linkage in Merging Cancer Registry Data and Electronic Records. Fall;52(3):79-83

Jain N, Carey D, Ham L, Steffens J, Hernández J, Zachary I. Modernizing Registry Operations: Implementing File*Pro Automation with Groovy Scripting to Streamline Data Processing in Missouri. Winter;52(4):155-159

Hanson, Heidi A.

Shivanna A, Fitts A, Tschida J, Chandrashekar M, Hwang A, Hsu E, Gounley J, Hanson HA. Leveraging Large Language Models for Real-World Data Evidence: A Framework for Automated Treatment Extraction and Data Harmonization. Winter;52(4):138-145.

Hartley, Jeanie

Moore R, Hartley J, Maguire F, Sherman R. Investigation of High Number of Primary Tumors Reported in Central Sequence Number. Winter;52(4):169

Hernández, Joseline

Jain N, Carey D, Ham L, Steffens J, Hernández J, Zachary I. Modernizing Registry Operations: Implementing File*Pro Automation with Groovy Scripting to Streamline Data Processing in Missouri. Winter;52(4):155-159

Hofer, Brenda M.

Maguire FB, Vernon ASM, Hofer BM, Parikh-Patel A, Wun T, Hussain SK, Keegan THM. The Effect of Persistent Poverty on Cancer Outcomes in California. Winter;52(4):109-118.

Hsieh, Meichin

Maniscalco L, Mumphrey B, Hsieh M, Wu XC. Using eMaRC Lite to Streamline Pathology Report Reviews and Enhance Rapid-Case Ascertainment Studies. Summer;52(2):50

Hsu, Elizabeth

Shivanna A, Fitts A, Tschida J, Chandrashekar M, Hwang A, Hsu E, Gounley J, Hanson HA. Leveraging Large Language Models for Real-World Data Evidence: A Framework for Automated Treatment Extraction and Data Harmonization. Winter;52(4):138-145.

Hunter, Nathan

Bayakly R, McNamara C, Blankenship E, Hunter N, Gatta P. The Georgia Cancer Registry Database Improves Local Cancer Data Access. Fall;52(3):99

L

Lanese, Robert

Rose J, Hussain F, Sahoo SS, Lyytinen K, Song S, Menegay HJ, Lanese R, SG A, Lihua L, Tsui J, Beno M, Burus T, Koroukian SM. The Population Cancer Assessment and Surveillance Engine (PopCASE): An Emerging Population Cancer Data Platform. Winter;52(4):151-154

Lee, Sangwon

Jang B, Lee S, Park S. A Comparative Analysis of Survival Based on Body Mass Index at Diagnosis in Non-Small Cell Lung Cancer Patients. Fall;52(3):85-89

Lefante, Christina

Lefante C, Maniscalco L, Wu XC. Addressing the Challenge of Geospatial Accuracy in Cancer Registration: The Lousisiana Tumor Registry Experience. Winter;52(4):160-162

Leung, Kin

Dimacali MB, Sipin-Baliwas A, Leung K, Tsai KY, Wilson S. Quality Control Measures: The Value Visual Editing (VE) Brings to Los Angeles County. Fall;52(3):90-91

Liu, Lihua

Rose J, Hussain F, Sahoo SS, Lyytinen K, Song S, Menegay HJ, Lanese R, SG A, Lihua L, Tsui J, Beno M, Burus T, Koroukian SM. The Population Cancer Assessment and Surveillance Engine (PopCASE): An Emerging Population Cancer Data Platform. Winter;52(4):151-154

Lyamuya, Eligius F.

Pallangyo A, Mashauri O, Kalonge S, Amour M, Mremi A, Ngocho JS, Balandya E, Kwesigabo G, Kidenya B, Mshana SE, Lyamuya EF, Sunguya BF, Bartlett J, Mmbaga BT. Burden of HIV among Patients Undergoing Cancer Treatment: Analysis of Population Cancer Registry in Northern Tanzania. Summer;52(2)30-34

Lyytinen, Kalle

Rose J, Hussain F, Sahoo SS, Lyytinen K, Song S, Menegay HJ, Lanese R, SG A, Lihua L, Tsui J, Beno M, Burus T, Koroukian SM. The Population Cancer Assessment and Surveillance Engine (PopCASE): An Emerging Population Cancer Data Platform. Winter;52(4):151-154

M

Maguire, Frances B.

Maguire FB, Vernon ASM, Hofer BM, Parikh-Patel A, Wun T, Hussain SK, Keegan THM. The Effect of Persistent Poverty on Cancer Outcomes in California. Winter;52(4):109-118.

Moore R, Hartley J, Maguire F, Sherman R. Investigation of High Number of Primary Tumors Reported in Central Sequence Number. Winter;52(4):169

Maniscalco, Lauren

Maniscalco L, Mumphrey B, Hsieh M, Wu XC. Using eMaRC Lite to Streamline Pathology Report Reviews and Enhance Rapid-Case Ascertainment Studies. Summer;52(2):50

Lefante C, Maniscalco L, Wu XC. Addressing the Challenge of Geospatial Accuracy in Cancer Registration: The Lousisiana Tumor Registry Experience. Winter;52(4):160-162

Mashauri, Onstard

Pallangyo A, Mashauri O, Kalonge S, Amour M, Mremi A, Ngocho JS, Balandya E, Kwesigabo G, Kidenya B, Mshana SE, Lyamuya EF, Sunguya BF, Bartlett J, Mmbaga BT. Burden of HIV among Patients Undergoing Cancer Treatment: Analysis of Population Cancer Registry in Northern Tanzania. Summer;52(2)30-34

McNamara, Chrissy

Bayakly R, McNamara C, Blankenship E, Hunter N, Gatta P. The Georgia Cancer Registry Database Improves Local Cancer Data Access. Fall;52(3):99

Menegay, Harry J.

Rose J, Hussain F, Sahoo SS, Lyytinen K, Song S, Menegay HJ, Lanese R, SG A, Lihua L, Tsui J, Beno M, Burus T, Koroukian SM. The Population Cancer Assessment and Surveillance Engine (PopCASE): An Emerging Population Cancer Data Platform. Winter;52(4):151-154

Miller, Jacqueline

Chapman SA, Hailer L, Miller J. Workload and Staffing Study of Hospital Cancer Registries. Spring;52(1):16-21

Mmbaga, Blandina T.

Pallangyo A, Mashauri O, Kalonge S, Amour M, Mremi A, Ngocho JS, Balandya E, Kwesigabo G, Kidenya B, Mshana SE, Lyamuya EF, Sunguya BF, Bartlett J, Mmbaga BT. Burden of HIV among Patients Undergoing Cancer Treatment: Analysis of Population Cancer Registry in Northern Tanzania. Summer;52(2)30-34

Moore, Rebecca

Moore R, Hartley J, Maguire F, Sherman R. Investigation of High Number of Primary Tumors Reported in Central Sequence Number. Winter;52(4):169

Morales, Caroline B.

Morales CB, Paddock LE, Bleeker K, Stroup AM. The Burden of Cancers Associated with Hereditary Breast and Ovarian Cancer Syndrome, 2015-2019: A Population-based Cancer Registry Analysis. Winter;52(4):119-126

Mremi, Alex

Pallangyo A, Mashauri O, Kalonge S, Amour M, Mremi A, Ngocho JS, Balandya E, Kwesigabo G, Kidenya B, Mshana SE, Lyamuya EF, Sunguya BF, Bartlett J, Mmbaga BT. Burden of HIV among Patients Undergoing Cancer Treatment: Analysis of Population Cancer Registry in Northern Tanzania. Summer;52(2)30-34

S

T

Key Word Index—Vol. 52 (2025)

A

B

C

Title Index—Vol. 52 (2025)

A

A Comparative Analysis of Survival Based on Body Mass Index at Diagnosis in Non-Small Cell Lung Cancer Patients

Jang B, Lee S, Park S. A Comparative Analysis of Survival Based on Body Mass Index at Diagnosis in Non-Small Cell Lung Cancer Patients. Fall;52(3):85-89

Addressing the Challenge of Geospatial Accuracy in Cancer Registration: The Lousisiana Tumor Registry Experience

Lefante C, Maniscalco L, Wu XC. Addressing the Challenge of Geospatial Accuracy in Cancer Registration: The Lousisiana Tumor Registry Experience. Winter;52(4):160-162

B

Building a Cancer Registry from the Ground Up: a Kaiser Permanente Case Study

Baranggay JA, Jefferson DR. Building a Cancer Registry from the Ground Up: a Kaiser Permanente Case Study. Fall;52(3):92-96

Burden of HIV among Patients Undergoing Cancer Treatment: Analysis of Population Cancer Registry in Northern Tanzania

Pallangyo A, Mashauri O, Kalonge S, Amour M, Mremi A, Ngocho JS, Balandya E, Kwesigabo G, Kidenya B, Mshana SE, Lyamuya EF, Sunguya BF, Bartlett J, Mmbaga BT. Burden of HIV among Patients Undergoing Cancer Treatment: Analysis of Population Cancer Registry in Northern Tanzania. Summer;52(2)30-34

C

CaRDO: A Tool for Empowering Cancer Registries to Publish Population-Level Cancer Statistics Online

Retell J, Francis S, Rene J, Baade P. CaRDO: A Tool for Empowering Cancer Registries to Publish Population-Level Cancer Statistics Online. Winter;52(4):146-150.

Chronological Trends of PSA Levels at Diagnosis in Population-based Prostate Cancer Data 2004-2021, a Time Period of Drastic Changes in Screening Guidelines: Potential Implications in Prostate Cancer Diagnosis

Kato I, Bosland MC, Kim S, Gann PH. Chronological Trends of PSA Levels at Diagnosis in Population-based Prostate Cancer Data 2004-2021, a Time Period of Drastic Changes in Screening Guidelines: Potential Implications in Prostate Cancer Diagnosis. Fall;52(3):70-78

Coalition Collaboration to Improve Timely Cancer Registry Data Reporting

Doyle D. Coalition Collaboration to Improve Timely Cancer Registry Data Reporting. Summer;52(2):49

D

Data Modernization Approach Reinvigorates Urban Reporting in Massachusetts

Knowlton R, Cherala S, Klein N, Pavlova-Plotnik R, Dooley D. Data Modernization Approach Reinvigorates Urban Reporting in Massachusetts. Fall;52(3):100

Data-Related Processes and Challenges at Level I–IV Trauma Centers in Arkansas: Results of a Mixed-Methods Case Study

Conner CL, Redding MC, Seker E, Greer ML, Williams TB, Garza MY. Data-Related Processes and Challenges at Level I–IV Trauma Centers in Arkansas: Results of a Mixed-Methods Case Study. Spring;52(1):6-15

E

Evaluating Privacy-Preserving Record Linkage in Merging Cancer Registry Data and Electronic Records

Wilcox A, Grannis S, Zachary I, Venterisv N, Abrams Z, Newland J, Beheshti M, Ham L, Day J, Barnholtz-Sloan JS. Evaluating Privacy-Preserving Record Linkage in Merging Cancer Registry Data and Electronic Records. Fall;52(3):79-83

F

Fall 2025 Continuing Education Quiz

Vida, C. Fall 2025 Continuing Education Quiz. Fall;52(3):23

I

Improving Cancer Reporting Compliance Across Maryland

Adamson T. Improving Cancer Reporting Compliance Across Maryland. Summer;52(2):52

Improving Reporting Timeliness Compliance in Oklahoma

Carter M, Dabbs C. Improving Reporting Timeliness Compliance in Oklahoma. Summer;52(2):56

Incorporating 2007-2022 Veterans Administration Data into the Cancer Registry in Less than One Month

Rogers D. Incorporating 2007-2022 Veterans Adminstration Data into the Cancer Registry in Less than One Month. Summer;52(2):53

Inferring Unknown Race in Central Cancer Registries

Boscoe FP. Inferring Unknown Race in Central Cancer Registries. Summer;52(2)42-47

Integrating Cancer Registry and Medicaid Data for Survivorship-Focused Research

Koroukian SM, Shoag J, Vu L, Rose J. Integrating Cancer Registry and Medicaid Data for Survivorship-Focused Research. Winter;52(4):163-167

Interactive Visualization of Cancer Data

Ramaswami J. Interactive Visualization of Cancer Data. Fall;52(3):101

National Cancer Registrars Association
CALL FOR PAPERS

The *Journal of Registry Management*, official journal of the National Cancer Registrars Association (NCRA), announces a call for original manuscripts on registry methodology or research findings related to the 7 subjects listed below and related topics.

Topics:

1. **Birth Defects Registries**
2. **Cancer Registries**
   a. **AJCC TNM Stage**
   b. **Cancer and Socioeconomic Status**
   c. **Cancer and Health Disparities**
3. **Trauma Registries**
4. **Recruitment, Training, and Retention**
5. **Public Relations**
6. **Quality Review**
7. **Registry Management**

Contributed manuscripts are peer-reviewed prior to publication. Manuscripts of the following types may be submitted for publication:

1. **Methodology Articles** addressing topics of broad interest and appeal to the readership, including methodological aspects of registry organization and operation.
2. **Research articles** reporting findings of original, reviewed, data-based research.
3. **Primers** providing basic and comprehensive tutorials on relevant subjects.
4. **"How I Do It" Articles** describe tips, techniques, or procedures for an aspect of registry operations that the author does particularly well. The "*How I Do It*" feature in the *Journal* provides registrars with an informal forum for sharing strategies with colleagues in all types of registries.
5. **Opinion papers/editorials** including position papers, commentaries, essays, and interviews that analyze current or controversial issues and provide creative, reflective treatments of topics related to registry management.
6. **Bibliographies** which are specifically targeted and of significant interest will be considered.
7. **Letters to the Editor** are also invited.

Address all manuscripts to: Nadine Walker, MS, ODS-C, Editor-in-Chief, *Journal of Registry Management,*
(703) 299-6640 ext. 327, JRMEditor@ncra-usa.org.

Manuscript submission requirements are given in "Information for Authors" found near the back of each *Journal* and on the NCRA website at **https://www.ncra-usa.org/About/Publications/Journal-of-Registry-Management**.

INFORMATION FOR AUTHORS

The Journal of Registry Management, the official journal of the National Cancer Registrars Association (NCRA), invites submission of original manuscripts on topics related to management of disease registries and the collection, management, and use of cancer, trauma, birth defects, HIV/AIDS, and other disease registry data. JRM is a peer-reviewed, open-access, online-only journal and is published quarterly.

JRM encourages authorship by registrars who are ODS-certified (Oncology Data Specialist); special value is placed on manuscripts with ODS-certified professionals' collaboration and publication of articles or documents related to the registry profession. Three NCRA continuing education (CE) credits are awarded for published articles or documents, and additional information can be found at the following URL: **https://www.ncra-usa.org/ODS-Credential/Current-ODS/Submit-CEs/CE-Eligible-Activities**

Manuscripts may be submitted for publication in the following categories:

- Articles addressing topics of broad interest and appeal to the readership, including methodology papers about registry organization and operation
- Research papers reporting findings of original research, literature reviews, data-based research
- Primers providing tutorials on relevant subjects
- "How I Do It" papers

JRM invites submission of:

- Opinion papers/editorials including position papers, commentaries, and essays that analyze current or controversial issues and provide creative, reflective treatments of topics related to registry management
- Letters to the Editor
- Specifically targeted bibliographies of significant interest

Previously published material will be considered for publication only when it is of special and immediate interest to the readership.

Please submit manuscripts and articles here: https://srvy.pro/2CXB3FV.

Manuscript Preparation Guidelines

The following guidelines are provided to help prospective authors prepare manuscripts for the JRM and facilitate technical processing of submissions. Failure to follow the guidelines may delay consideration of your manuscript.

Authors who are unfamiliar with preparation and submission of manuscripts for publication are encouraged to contact the Editor for clarification or additional assistance. All correspondence and questions about manuscripts should be sent to **JRMEditor@ncra-usa.org**. Telephone inquiries may be directed to (703) 299-6640 ext. 327

Cover Letter and Signature Page

An accompanying cover letter should include the name, mailing address, email address, and telephone number of the corresponding author(s).

An authors' signature page can be a scanned copy containing all the author's signatures, or an email acknowledgement from each author can be sent to **JRMEditor@ncra-usa.org**. See Copyright section below for instructions on authors' permissions.

Manuscript Types

The terms *manuscripts, articles,* and *papers* are used synonymously herein. Number the manuscript pages consecutively with the title page as page 1, followed by the abstract, text, references, and visuals.

Articles

Articles should follow the standard format for research reporting (Introduction, Methods, Results, Discussion, References). The introduction will normally include background information, and a rationale/justification as to why the subject matter is of interest. The discussion often includes a conclusion subsection. Comprehensive references are encouraged, as are an appropriate combination of tables and figures (graphs).

Methodology/Process Papers

Methodology papers should follow the standard format for research reporting (Introduction, Methods, Results, Discussion), or for explanatory papers not reporting results (Introduction, Methods, Discussion).

"How I Do It" Articles

The "How I Do It" feature in the *JRM* provides registrars with a forum for sharing strategies with colleagues in all types of registries. These articles describe tips, techniques, or procedures for an aspect of registry operations that the author does particularly well. When shared, these innovations can help registry professionals improve their skills, enhance registry operations, or increase efficiency.

"How I Do It" articles should be 1,500 words or less (excluding references) and can contain up to 2 tables or figures. To the extent possible, the standard headings (Introduction, Methods, Results, Discussion) should be used. If results are not presented, that section may be omitted. Authors should describe the problem or issue, their solution, advantages (and disadvantages) to the suggested approach, and their conclusion. All submitted "How I Do It" articles will have the benefit of peer/editorial review.

Manuscript Format and Structure

Authors

Each author 's name, degrees, certifications, title, professional affiliation, and email address must be noted on the title page exactly as it is to appear in publication. The corresponding author should be noted, with mailing address included. Joint authors should be listed in order of their contribution to the work. A maximum of 6 authors for each article will be listed. Consider using a "working group" or committee name if there are more than 6 authors and list the lead author(s).

Title

Authors are urged to choose a title that accurately describes the manuscript's content. Every effort will be made to use the title as submitted; however, *JRM* reserves the right to select a title that is consistent with editorial and production requirements.

Abstract

A brief abstract must accompany each article or manuscript (word limit, 350). The abstract should summarize the main point(s) and quickly give the reader an understanding of the manuscript's content. It should be placed on a page by itself, immediately following the title page.

Key words

Authors are asked to provide up to 5 alphabetized key words or phrases which will be used in compiling the Annual Subject Index. Key words should be included directly under the abstract on the abstract stand-alone page.

Length

The word count for manuscripts should not exceed 3,000 words. Word count guidelines exclude abstracts, figure legends, and table notes. Authors are invited to contact the Editor regarding submissions of markedly longer manuscripts.

Style

Prepare manuscripts using the *American Medical Association Manual of Style,* 11th edition (2020). All sections of the paper should be single-spaced. Double-space between paragraphs and sections.

Visuals

Use visuals selectively to supplement the text. Visual elements—charts, graphs, tables, diagrams, and figures—will be reproduced exactly as received. Copies must be clear and properly identified. Each visual must have a brief, self-explanatory title. Submit each visual on a separately numbered page at the end of the manuscript, following the references. Visuals for research papers, articles, and methodology/process papers are limited to 5; Visuals for "How I Do It" papers and all other submissions are limited to 2.

Attribution

Authors are to provide appropriate acknowledgment of products, activities, and support, especially for those articles based on or utilizing registry data (including acknowledgment of hospital and central registrars). Appropriate attribution is also to be provided to acknowledge federal funding sources of registries from which the data are obtained.

References

References should be carefully selected, and relevant. References must be numbered in order of their appearance in the text. At the end of the manuscript, list the references in the order they are cited; do not list references alphabetically. Journal citations should include author, title, journal, year, volume, issue, and pages. Book citations should include author, title, city, publisher, year, and pages. Authors are responsible for the accuracy of all references.

Examples:

1. LeMaster PL, Connell CM. Health education interventions among Native Americans: a review and analysis. *Health Educ.* 1995;21(4):521-538. doi:10.1177/109019819402100413
2. Hanks GE, Myers CE, Scardino PT. Cancer of the prostate. In: DeVita VT, Hellman S, Rosenberg SA. *Cancer: Principles and Practice of Oncology.* 4th ed. J.B. Lippincott Co.; 1993:1073-1113.

Affirmations

Copyright

Authors submitting a manuscript do so on the understanding that if it is accepted for publication, copyright of the article, including the right to reproduce the article in all forms and media, shall be assigned exclusively to NCRA. NCRA will not refuse any reasonable requests by the author(s) for permission to reproduce any of his or her contributions to the *JRM.*

Authors Permissions

Further, the manuscript's accompanying cover letter, signed by all authors, must include the following statement: "We, the undersigned, transfer to the National Cancer Registrars Association, the copyright for this manuscript in the event that it is published in *Journal of Registry Management*." Failure to provide the statement will delay consideration of the manuscript. The authors' signature page can be a scanned copy containing all corresponding authors signature, or an email acknowledgement from each author can be sent to JRMEditor@ncra-usa.org. It is the author's responsibility to obtain necessary permission when using material (including graphs, charts, pictures, etc) that has appeared in other published works.

Originality

Articles are reviewed for publication assuming they have not been accepted or published previously and are not under simultaneous consideration for publication elsewhere. If the article has been previously published or significantly distributed, this is to be noted in the cover letter for consideration.

Editing

Journal of Registry Management reserves the right to edit all contributions for clarity and length. Minor changes (punctuation, spelling, grammar, syntax) will be made at the editorial staff's discretion. Substantive changes will be verified with the author(s) prior to publication.

Peer Review

JRM follows a double-blind peer review process. Contributed manuscripts are reviewed prior to publication, generally by 2 reviewers. The Journal Editor makes the final decision regarding acceptance of manuscripts. Receipt of manuscripts will be acknowledged promptly, and corresponding authors will be advised of the status of their submission as soon as possible.

Ethics

Conflict of Interest

As part of the online submission process, corresponding authors must confirm if they or their coauthors have conflicts of interest to declare, and to provide details of these. These include all financial and non-financial interests and relationships, direct or indirect, or other situations that might raise questions of bias in the work reported or the conclusions, implications, or opinions stated. Authors should also disclose any conflict of interest that may have influenced either the conduct or the presentation of the research to the editors, including but not limited to, close relationships with those who might be helped or hurt by the publication, academic interests, and rivalries, and any personal, religious, or political convictions relevant to the topic at hand. If the manuscript is published, conflict of interest information will be communicated in a statement within the published paper. If any reviewer believes there is likely to be a perception of a conflict of interest in relation to their review of a submitted manuscript, they will notify the Editor-in-Chief. The review will then be assigned to other Editorial Advisory Board (EAB) members. Authors should identify individuals who provide writing assistance and disclose the funding source for this assistance.

Informed Consent

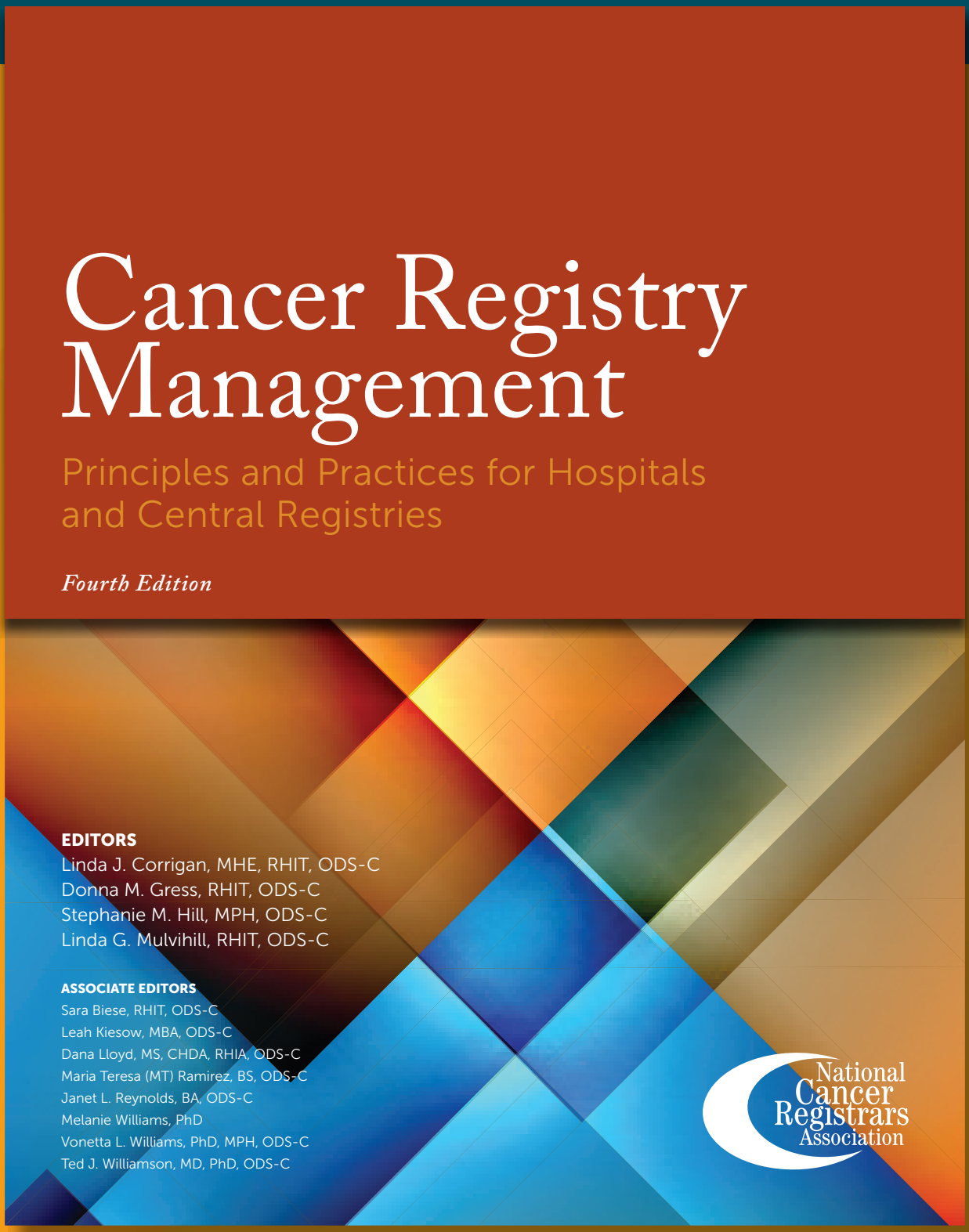
Individuals have a right to privacy that should not be infringed without informed consent. Identifying information, including patients' names, initials, or hospital numbers, should not be published in written descriptions, photographs, and pedigrees, unless the information is essential for scientific purposes and the patient (or parent or guardian) gives written informed consent for publication. Informed consent for this purpose requires that a patient who is identifiable be shown the manuscript to be published. Identifying details should be omitted if they are not essential. Complete anonymity is difficult to achieve; however, an informed consent should be obtained if there is any doubt. For example, masking the eye region in photographs of patients is inadequate protection of anonymity. If identifying characteristics are altered to protect anonymity, such as in genetic pedigrees, authors should provide assurance that alterations do not distort scientific meaning and editors should so note.

Human and Animal Rights

Research involving human subjects, including identifiable human material or data, must have been performed in accordance with the Declaration of Helsinki and must have been approved by an appropriate ethics committee. A statement detailing this, including the name of the ethics committee and the reference number where appropriate, must appear in all manuscripts reporting such research. Although *JRM* does not publish animal research, it may accept it in specific situations (eg, when an animal experiment is also part of a human trial). Authors interested in submitting animal research should contact the Editor-in-Chief. Any study using animals needs to state the Institutional Animal Care approval and number. Any other ethics approvals should also be listed. If no ethical approvals were required, please state this.

A definitive update of the premier textbook and desk reference for the cancer registry profession.

Order print or e-book at
www.ncra-usa.org/4thEdition *today!*



Journal of Registry Management

NCRA Executive Office
1330 Braddock Place
#520
Alexandria, VA 22314
www.ncra-usa.org
JRMEditor@ncra-usa.org