





RESEARCH BRIEF

The role of all-payer claims databases to expand central cancer registries: Experience from Colorado

Marcelo C. Perrailon PhD¹  | Rifei Liang MS²  | Lindsay M. Sabik PhD³  |
Richard C. Lindrooth PhD¹ | Cathy J. Bradley PhD⁴ 

¹Health Systems, Management & Policy, University of Colorado—Anschutz Medical Campus, Aurora, Colorado, USA

²University of Colorado Cancer Center, University of Colorado—Anschutz Medical Campus, Aurora, Colorado, USA

³Health Policy and Management, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

⁴Colorado School of Public Health, University of Colorado, Aurora, Colorado, USA

Correspondence

Marcelo C. Perrailon, Health Systems, Management & Policy, University of Colorado—Anschutz Medical Campus, 13001 E. 17th Place, E3312 Aurora, CO 80045, USA.
Email: marcelo.perrailon@cuanschutz.edu

Funding information

National Cancer Institute, Grant/Award Numbers: P30CA046934, R01CA22599

Abstract

Objective: To evaluate the quality of a multiyear linkage between the Colorado all-payer claims database (APCD) and the Colorado Central Cancer Registry.

Data sources: Secondary 2012–2017 data from the APCD and the Colorado Cancer Registry.

Study design: Descriptive analysis of the proportion of cases captured by the linkage in relation to the cases reported by the registry.

Data collection/extraction methods: We used probabilistic linkage to combine records from both data sources for all patients diagnosed with cancer.

Results: We successfully linked 93% of the 146,884 patients in the registry. Approximately 63% of linked patients were perfect matches on five identifiers. Of partial matches, 81.6% were matched on four identifiers with missing or partial Social Security Numbers. The linkage rate was lower for uninsured patients at diagnosis (74.7%) or patients with private plans (89.4%) but close to 100% for Medicare and Medicaid enrollees. Most of the 29% of patients who did not have claims at the time of diagnosis were covered by private plans that may not submit claims.

Conclusions: APCD-registry linkages are a promising source of data to conduct population-based research from multiple payers. However, not all payers submit claims, and the quality of the data may vary by state.

KEYWORDS

all-payer claims databases, cancer, linkage, longitudinal, registry

What is known on this topic

- Health services research on cancer care and outcomes has benefited from linkages between claims databases and cancer registries, but these linkages are limited to a single payer, health system, or population.
- Many states have developed all-payer claims databases, but each state has different rules regarding claims submission and not all payers are required to submit claims.
- The quality of all-payer claims databases for cancer health services research is unknown.

What this study adds

- We linked most patients (93.0%) in the Colorado Cancer Registry with the Colorado all-payer claims databases, with high linkage rates of 98.6% for Medicare and 99.2% for Medicaid, and similar linkage rates by race and urban/rural residence (89.3%–94.8%).

- Although we were able to link 93% of patients in the registry, nearly 30% of patients did not have claims at the time of diagnosis, either due to uninsurance or coverage from private plans that do not submit claims.
- Even though nearly 30% of linked patients did not have claims data at the time of diagnosis, claims available in the linkage after diagnosis could be valuable in some study designs (e.g., survivorship studies) and baseline characteristics for these patients are available from the registry.

1 | INTRODUCTION

Cancer health services research has benefited from the availability of claims data linked to cancer registry data; the most well-known example is Surveillance, Epidemiology, and End Results (SEER)-Medicare.¹ Claims include patient-level longitudinal information on cancer screening, treatment, and payments. However, claims data alone are limited. Precise diagnosis date, cancer stage, tumor characteristics, and vital status are not present. In contrast, data collected through cancer registries have excellent patient- and tumor-level diagnosis date and stage data. They also capture race and ethnicity but may not adequately capture treatment information beyond the first course of treatment.² Linkage of claims to registry data can significantly expand the capability of each source. To date, most linkages are performed using data from a single-payer, such as Medicare,^{3,4} Medicaid,^{5,6} or a handful of private payers.^{7,8} Some registries have linked to statewide hospital discharge databases,^{9,10} covering all payers, including uncompensated care, but these data are pertinent only to inpatient care.

All-payer claims databases (APCDs) serve as a depository for public and private claims for health care services provided to insured individuals within a state.¹¹ Twenty-one states have established APCDs, and many others are currently in development.^{12,13} APCDs linked to cancer registry data can potentially provide longitudinal data to study cancer care and outcomes among insured individuals across multiple payers. This information can be used for evaluating the effects of coverage disruptions and differences in state and insurer-level policies, such as Medicaid generosity, changes in reimbursement policies, managed care, and the care of dually eligible patients (Medicare and Medicaid).¹⁴ However, as of 2021, only one state, Utah, reported an APCD-registry linkage and only for 1 year of data.¹⁵ Recent reports by The Commonwealth Fund found that states have implemented APCDs in diverse ways, from governmental initiatives to public-private partnerships and voluntary efforts. Consequently, the governance of APCDs, their funding, and the authority of APCDs to require claims submissions also varies by state.¹⁶ The extent to which APCD data covers those with a cancer diagnosis, and the variability of data quality by state is not known.

This article aims to evaluate the quality and comprehensiveness of a multiyear linkage between the Colorado APCD and the Colorado Central Cancer Registry (CCCR). We evaluate the quality of the probabilistic linkage and the proportion of cases captured by the linkage in relation to the cases reported by the registry. We also explore the

implication of incomplete data as not all payers must submit claims to APCDs.

2 | METHODS

2.1 | Data

The Center for Improving Value in Health Care (CIVHC), a nonprofit organization, was authorized by the state to collect Colorado claims data. The APCD includes medical claims and dates of service from commercial health plans, Medicare, and Colorado's Medicaid Program. Plans offered by self-insured employers are not required to report, but some do so voluntarily. These plans are regulated under the Employee Retirement Income Security Act (ERISA). Private payers covering fewer than 100 enrolled individuals are also not required to submit claims.

The CCCR includes cancer site, stage of disease at diagnosis, month and year of diagnosis, initial treatment, insurance, and demographic characteristics, including age, sex, race, Hispanic ethnicity, marital status, and county of residence, which is the smallest geographical unit released by the CCCR. However, under our Data Use Agreement, the CCCR dataset included variables that were coded using smaller geographical units, which allow us to code Geographic Underserved Areas as defined by the National Cancer Institute (NCI).¹⁷ At the Census tract level, these variables included the percent of the population living at or below 100% of the Federal poverty level and the percent of the population with only high school education or less, both obtained from the American Community Survey.¹⁸ Variables defined at the county level include indicators of whether a person resides in a Rural-Urban Commuting Area (RUCA),¹⁹ a Health Professional Shortage Area (HPSA),²⁰ nonmetro area,²¹ a high-poverty area, or a persistent poverty area.²²

2.2 | Linkage methodology

According to Colorado statutes, the CCCR cannot release identifying information. Therefore, the CCCR conducted the linkage and deidentified the data before releasing them to the research team. CIVHC sent identifiers (APCD member ID, Social Security Number (SSN), date of birth (DOB), last name, first name, and sex) of all individuals older than 21 who appeared at least once in the APCD from

2012 to 2017. The CCCR registrar sent APCD member IDs of patients who were successfully linked. CIVHC then extracted all claims for linked individuals and sent them to the investigators. The CCCR sent investigators deidentified information for patients linked and not linked.

The linkage was performed using a probabilistic linkage approach following the Fellegi-Sunter model^{23,24} as implemented by Match*Pro.²⁵ We used the following five data fields: SSN, DOB, last name, first name, and sex. Match*Pro reports a score for each of the fields based on an assessment of the similarity between data fields. The total score is the sum of the scores for each of the five fields. Partial similarity in key fields (e.g., transposition of digits in SSN) was incorporated. If a particular data field is not a perfect match, the score for that field is lowered or could be negative if too dissimilar, reducing the overall score. If a data field is missing, the score for that data field is set to zero. Matched sets of individuals with scores lower than 15 were considered unsuccessful matches. Each field score is helpful for review since similar values (i.e., consecutive SSN numbers) would score higher than larger differences (or missing values). Matched sets of individuals with scores greater or equal than 15 were considered potentially linked records and further evaluated.

The exact matches, defined as pairs that matched all five data fields, were accepted as true linked pair. Partial matches required a more detailed manual review. The strategy was first to review records that matched on four out of five identifiers. For example, a pair may not have the same DOB, but if it matched on all the other identifiers and the month and day of birth were reversed, it is very likely that it is a true match. Similar decision rules were applied to partial or transposed SSN digits, maiden names, transposed last names (more common in Latino individuals who use two last names without a hyphen), and misspelled first or last names. The next step considered partial matches of three out of five identifiers and then two out of five. The only identifier that could be used by itself was SSN, although at least one of the other four fields had to provide partial confirmation. The manual review continued until each pair was classified as linked or not linked.

2.3 | Linkage validation

To evaluate the quality of the linkage, we treated the CCCR data as the gold standard. We defined the linkage rate as the number of patients who were linked across the datasets divided by the total number of patients in the registry. In addition, we compared the characteristics of linked and nonlinked patients based on demographic characteristics, payer, site, stage at diagnosis, census tract poverty level and education, and county-level information. Because of large sample sizes, even small differences are statistically significant. We report Chi-square *p* values in footnotes when they are greater than 0.05. We excluded patients with missing demographic characteristics, missing payer information, and those with no information on tumor characteristics.

Because CIVHC sent CCCR information for any person who was ever present in the APCD from 2012 to 2017, we expected a high linkage rate because of the length of time available to find a match. However, a high linkage rate does not imply that claims are available at the time of diagnosis. For example, patients could be linked because they were enrolled in Medicare or Medicaid at some point during 2012–2017, but they were uninsured or were enrolled in a private plan that does not submit claims at the time of diagnosis. For these types of linkages to be useful for health services research, it is important to evaluate the completeness of the data at the time of cancer diagnosis. Therefore, we present characteristics for patients with and without an APCD health plan at the time of diagnosis. To comply with CCCR privacy regulations, the linked dataset has the month and year of diagnosis, not the exact date. We defined time of diagnosis based on a window starting one calendar month before the month of diagnosis and three calendar months after the month of diagnosis.

The University of Colorado Institutional Review Board reviewed and approved this study.

TABLE 1 Linkage statistics by final link status, 2012–2017

Total	Number of patients	Percent
Linked—exact match on five identifiers	91,028	62.90
Linked—partial match (less than five identifiers)	45,585	31.50
Not linked—partial match (less than five identifiers)	8114	5.61
Total	144,727	100
A. Linked—partial match (N = 45,585)		
DOB, first name, last name, sex, missing SSN	32,609	71.53
DOB, first name, last name, sex, partial SSN	4603	10.10
SSN complete or partial only	5882	12.90
DOB, first name, last name only	398	0.87
Other combinations	2093	4.59
Total	45,585	100
B. Not linked—partial match (N = 8114)		
DOB complete only	3300	40.67
First name, sex only	2442	30.10
Last name only	657	8.10
Last name, sex only	635	7.83
SSN partial only	507	6.25
Other	573	7.07
Total	8114	100

Note: Partial SSN refers to SSN with less than nine digits. Partial match means that individuals were potential matches in less than five data fields. Panel A shows the most common combinations that resulted in linkages after review even though they were not perfect matches. Panel B shows the most common reasons potential matches were not linked. Abbreviations: DOB, date of birth; SSN, social security number.

TABLE 2 Characteristics of linked and nonlinked patients, 2012–2017

Characteristic based on registry	Not linked (N = 10,271)	Linked (N = 136,613)	Linkage rate (%)
Year of diagnosis			
2012	2113 (20.6)	22,129 (16.2)	91.28
2013	1789 (17.4)	22,331 (16.3)	92.58
2014	1597 (15.5)	22,810 (16.7)	93.46
2015	1593 (15.5)	22,902 (16.8)	93.50
2016	1606 (15.6)	23,114 (16.9)	93.50
2017	1573 (15.3)	23,327 (17.1)	93.68
Primary payer at diagnosis			
Uninsured	813 (7.9)	2396 (1.8)	74.67
Private insurance	5112 (49.8)	43,234 (31.6)	89.43
Medicaid	84 (0.8)	10,395 (7.6)	99.20
Medicare	814 (7.9)	56,207 (41.1)	98.57
Dual Medicare-Medicaid	17 (0.2)	3325 (2.4)	99.49
Other ^a	3431 (33.4)	21,056 (15.4)	85.99
Sex^b			
Female	5488 (53.4)	71,732 (52.5)	92.89
Male	4783 (46.6)	64,881 (47.5)	93.13
Age category			
21–40	1281 (12.5)	8439 (6.2)	86.82
41–60	6211 (60.5)	39,514 (28.9)	86.42
61–80	2316 (22.5)	71,064 (52.0)	96.84
Over 80	463 (4.5)	17,596 (12.9)	97.44
Race/ethnicity category			
White/non-Hispanic	7901 (76.9)	112,080 (82.0)	93.41
White/Hispanic	1309 (12.7)	13,962 (10.2)	91.43
Black	390 (3.8)	4535 (3.3)	92.08
Other	341 (3.3)	3280 (2.4)	90.58
Unknown	330 (3.2)	2756 (2.0)	89.31
Marital status			
Missing	574 (5.6)	7316 (5.4)	92.72
Not married or partnered	3482 (33.9)	54,609 (40.0)	94.01
Married or partnered	6215 (60.5)	74,688 (54.7)	92.32
Patient rural/urban commuting area^c			
Missing	828 (8.1)	8864 (6.5)	91.46
Not an urban commuting area	1158 (11.3)	15,922 (11.7)	93.22
Urban commuting area	8285 (80.7)	111,827 (81.9)	93.10
Primary site			
Breast	1927 (18.8)	24,552 (18.0)	92.72
Prostate	894 (8.7)	15,033 (11.0)	94.39
Lung	822 (8.0)	12,239 (9.0)	93.71
Melanoma	794 (7.7)	10,960 (8.0)	93.24
Colorectal	704 (6.9)	10,262 (7.5)	93.58
Brain and other nervous system	503 (4.9)	5991 (4.4)	92.25
Lymphoma	392 (3.8)	5601 (4.1)	93.46
Other	4235 (41.2)	51,975 (38.0)	92.47

TABLE 2 (Continued)

Characteristic based on registry	Not linked (N = 10,271)	Linked (N = 136,613)	Linkage rate (%)
SEER summary stage^d			
In situ	887 (8.6)	14,199 (10.4)	94.12
Localized	3847 (37.5)	55,933 (40.9)	93.56
Regional	1964 (19.1)	24,854 (18.2)	92.68
Distant	2286 (22.3)	28,392 (20.8)	92.55
N/A or unstaged	1287 (12.5)	13,235 (9.7)	91.14
Census tract poverty level^e			
Missing	844 (8.2)	8881 (6.5)	91.32
Less than median	5234 (51.0)	65,880 (48.2)	92.64
Median or higher	4193 (40.8)	61,852 (45.3)	93.65
Census tract % HS or less univ. 25+^f			
Missing	828 (8.1)	8865 (6.5)	91.46
Less than median	5196 (50.6)	64,844 (47.5)	92.58
Median or higher	4247 (41.3)	62,904 (46.0)	93.68
HPSA^g			
No	10,228 (99.6)	135,730 (99.4)	92.99
Yes	43 (0.4)	883 (0.6)	95.36
Nonmetro area^h			
No	10,149 (98.8)	134,381 (98.4)	92.98
Yes	122 (1.2)	2232 (1.6)	94.82
High poverty areaⁱ			
No	10,149 (98.8)	134,381 (98.4)	92.98
Yes	122 (1.2)	2232 (1.6)	94.82
Persistent poverty area^j			
No	10,227 (99.6)	135,969 (99.5)	93.00
Yes	44 (0.4)	644 (0.5)	93.60

Note: Data presented as number of patients and (percentage). Chi-squared tests *p* values for sex, patient rural/urban commuting area, and primary site are greater than 0.05.

Abbreviations: HPSA, Health Professional Shortage Area; SEER, Surveillance, Epidemiology, and End Result.

^aOther (N = 24,487) includes registry categories Insurance Not Otherwise Specified (NOS) (3569), TRICARE (2365), Military (141), Veterans Affairs (3028), Indian/Public Health Service (26), Insurance status Unknown (12,759), and missings (2599).

^bEighteen cases have missing or unknown information.

^cClassification of Rural–Urban Commuting Areas based on 2010 Rural–Urban Commuting Area Codes using county FIPS codes (<https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/>).

^dSEER summary stage grouped by SEER Summary Staging Manual 2000.

^eCensus tract variable from American Community Survey (ACS) 5-Year data income and poverty for poverty status (using 100% level) in the past 12 months of families.

^fCensus Tract variable from American Community Survey (ACS) 5-year data educational attainment for the population 25 years and over.

^gFor shortage assignment of Health Professional Shortage Area. <https://bhw.hrsa.gov/shortage-designation/hpsas>. Geographic lever is county.

^h<https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/#map>.

ⁱ<https://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being/>.

^jBased on Geographically Underserved Areas: <https://cancercontrol.cancer.gov/research-emphasis/underserved.html>.

3 | RESULTS

The linkage was performed using data for 146,884 patients with a first diagnosed cancer between 2012 and 2017. Of these patients, 2157 (1.5%) had a linkage score below 15 and were not considered in the manual review. Table 1 reports linkage statistics for the remaining

144,727 patients. Of these patients, 5.61% were not successfully linked after manual review. Close to 63% of linked patients were perfect matches on five identifiers. Of the partial matches, most (81.6%) were matched because of an exact match on all identifiers except SSN. In these cases, SSN could be missing or had partial digits. In some cases (12.9%), the match was successful because the only

TABLE 3 Linked patients by whether an eligible plan is recorded in the Colorado all-payer claims databases at the time of diagnosis, 2012–2017^a

Characteristic based on registry	With APCD plan at diagnosis (N = 96,721)	Without APCD plan at diagnosis (N = 39,892)	Linkage rate at diagnosis (%)
Year of diagnosis			
2012	14,623 (15.1)	7506 (18.8)	60.32
2013	15,213 (15.7)	7118 (17.8)	63.07
2014	16,692 (17.3)	6118 (15.3)	68.39
2015	16,718 (17.3)	6184 (15.5)	68.25
2016	16,901 (17.5)	6213 (15.6)	68.37
2017	16,574 (17.1)	6753 (16.9)	66.56
Primary payer at diagnosis			
Uninsured	1182 (1.2)	1214 (3.0)	36.83
Private insurance	20,111 (20.8)	23,123 (58.0)	41.60
Medicaid	9894 (10.2)	501 (1.3)	94.42
Medicare	47,926 (49.6)	8281 (20.8)	84.05
Dual Medicare-Medicaid	3191 (3.3)	134 (0.3)	95.48
Other ^b	14,417 (14.9)	6639 (16.6)	58.88
Sex			
Female	49,722 (51.4)	22,010 (55.2)	64.39
Male	46,999 (48.6)	17,882 (44.8)	67.47
Age category			
21–40	4601 (4.8)	3838 (9.6)	47.34
41–60	21,050 (21.8)	18,464 (46.3)	46.04
61–80	55,034 (56.9)	16,030 (40.2)	75.00
Over 80	16,036 (16.6)	1560 (3.9)	88.80
Race/ethnicity category			
White/non-Hispanic	78,965 (81.6)	33,115 (83.0)	65.81
White/Hispanic	10,255 (10.6)	3707 (9.3)	67.15
Black	3309 (3.4)	1226 (3.1)	67.19
Other	2315 (2.4)	965 (2.4)	63.93
Unknown	1877 (1.9)	879 (2.2)	60.82
Marital status			
Missing data	5167 (5.3)	2149 (5.4)	65.49
Not married or partnered	42,335 (43.8)	12,274 (30.8)	72.88
Married or partnered	49,219 (50.9)	25,469 (63.8)	60.84
Patient rural/urban commuting area ^c			
Missing data	6506 (6.7)	2358 (5.9)	67.13
Not an urban commuting area	11,995 (12.4)	3927 (9.8)	70.23
Urban commuting area	78,220 (80.9)	33,607 (84.2)	65.12
Primary site			
Breast	15,678 (16.2)	8874 (22.2)	59.21
Prostate	10,088 (10.4)	4945 (12.4)	63.34
Lung	10,364 (10.7)	1875 (4.7)	79.35
Melanoma	7195 (7.4)	3765 (9.4)	61.21
Colorectal	7548 (7.8)	2714 (6.8)	68.83
Brain and other nervous system	4164 (4.3)	1827 (4.6)	64.12
Lymphoma	3996 (4.1)	1605 (4.0)	66.68
Other	37,688 (39.0)	14,287 (35.8)	67.05

TABLE 3 (Continued)

Characteristic based on registry	With APCD plan at diagnosis (N = 96,721)	Without APCD plan at diagnosis (N = 39,892)	Linkage rate at diagnosis (%)
SEER summary stage ^d			
In situ	9377 (9.7)	4822 (12.1)	62.16
Localized	37,623 (38.9)	18,310 (45.9)	62.94
Regional	17,120 (17.7)	7734 (19.4)	63.84
Distant	22,342 (23.1)	6050 (15.2)	72.83
N/A or unstaged	10,259 (10.6)	2976 (7.5)	70.64
Census tract poverty level ^e			
Missing data	6511 (6.7)	2370 (5.9)	66.95
Less than median	44,118 (45.6)	21,762 (54.6)	62.04
Median or higher	46,092 (47.7)	15,760 (39.5)	69.79
Census tract % HS or less university 25+ ^f			
Missing data	6506 (6.7)	2359 (5.9)	67.12
Less than median	43,201 (44.7)	21,643 (54.3)	61.68
Median or higher	47,014 (48.6)	15,890 (39.8)	70.01
HPSA ^g			
No	95,994 (99.2)	39,736 (99.6)	65.77
Yes	727 (0.8)	156 (0.4)	78.51
Nonmetro area ^h			
No	94,907 (98.1)	39,474 (99.0)	65.67
Yes	1814 (1.9)	418 (1.0)	77.06
High poverty area ⁱ			
No	94,907 (98.1)	39,474 (99.0)	65.67
Yes	1814 (1.9)	418 (1.0)	77.06
Persistent poverty area ^j			
No	96,211 (99.5)	39,758 (99.7)	65.81
Yes	510 (0.5)	134 (0.3)	74.13

Note: Data presented as number of patients and (percentage). No Chi-squared test *p* value is greater than 0.001. Linkage rate at diagnosis is Number of patients with APCD plan at diagnosis over number of patients in registry (Table 1) for each characteristic.

Abbreviations: APCD, all-payer claims database; HPSA, Health Professional Shortage Area; HS, high school; SEER, Surveillance, Epidemiology, and End Result.

^aDiagnosis date is only available in linked data as year and month. We defined “time of diagnosis” as 1 month before diagnosis and 3 months after.

^bOther (N = 21,056) includes registry categories Insurance Not Otherwise Specified (NOS) (3027), TRICARE (1343), Military (38), Veterans Affairs (2628), Indian/Public Health Service (23), Insurance status Unknown (11,613) and missing (2384).

^cClassification of Rural–Urban Commuting Areas based on 2010 Rural–Urban Commuting Area Codes using county FIPS codes (<https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/>).

^dSEER summary stage grouped by SEER Summary Staging Manual 2000.

^eCensus tract variable from American Community Survey (ACS) 5-year data income and poverty for poverty status (using 100% level) in the past 12 months of families.

^fCensus tract variable from American Community Survey (ACS) 5-year data educational attainment for the population 25 years and over.

^gFor shortage assignment of Health Professional Shortage Area, <https://bhwh.hrsa.gov/shortage-designation/hpsas>. Geographic lever is county.

^h<https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/#map>.

ⁱCounty-level poverty review <https://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being/>.

^jBased on Geographically Underserved Areas: <https://cancercontrol.cancer.gov/research-emphasis/underserved.html>.

information available was SSN. Most patients who were not linked had multiple fields missing.

Table 2 reports the characteristics of linked and not linked patients and linkage rates. The overall linkage rate was 93%. As expected, the linkage rate was lower (74.7%) for those who were

reported as uninsured at the time of diagnosis by the registry, those who were younger, and those who had private or other insurance coverage at the time of diagnosis (89.4%). In contrast, the linkage rate was almost 100% for patients insured by Medicaid, Medicare, or both (duals). The linkage rate was slightly lower in 2012 compared to later

years as reflected by the higher proportion of nonlinked patients in 2012. This was the first year of data CIVHC recommends using, which may be of slightly lower quality. The difference in the distribution of characteristics is statistically significant for most characteristics except sex, urban/rural residence, and persistent poverty. However, the magnitudes of the differences are small.

Table 3 compares those with and without a health plan in the APCD at the time of diagnosis. Of the 136,613 patients successfully linked, 29.2% did not have a plan in the APCD at the time of diagnosis. Most of these patients had a private insurance plan (58.0%), were uninsured (3.0%), or had other plans (16.6%). The "Other" category includes patients with insurance status as unknown, missing, and not otherwise specified (80.9%), Veteran Affairs (12.5%), and TRICARE (6.6%). Thus, the most likely explanation for missing information for these patients is that their insurance plan was not part of the APCD. Approximately, 14.6% of patients with Medicare coverage (according to the registry) did not have a plan in the APCD at the time of diagnosis. As the linkage statistics in Table 2 show, most of these patients were in the APCD at some point in 2012–2017. The majority (89.1%) are age 65 years or older, so it is likely that they are enrolled in Medicare. These patients could have been misclassified by the registry as having Medicare fee-for-service (FFS) but were in a Medicare Advantage plan administered by a private insurer that did not submit claims to the APCD. Our conclusions were unchanged when using a smaller time window around diagnosis.

4 | DISCUSSION

Our multiyear linkage of the CO APCD and CCCR resulted in a high linkage rate, with near-perfect rates for individuals enrolled in Medicare and Medicaid according to the registry. Close to two-thirds of linked patients were exact matches based on five identifiers that included SSN. Of the partial matches, over 80% were matched based on four identifiers (first name, last name, DOB, and sex) but had either a missing or partial SSN. These statistics provide reassurance that linked records are true matches. The linkage rate was lower for those who were labeled as uninsured by the registry or had a private plan that did not submit claims to the APCD.

Because we conducted the linkage using multiple years, we linked patients who had an APCD plan at any point during the period. When we restricted the analysis to a time window around diagnosis, close to 30% of the linked individuals did not have a health plan in the APCD, resulting in a lower linkage rate at diagnosis. However, claims available after diagnosis could still be used for research, and their baseline information is available from the registry. The most likely cause is that these individuals were in plans that do not submit claims data or were uninsured at the time of diagnosis. This highlights some of the concerns about APCDs. Despite their name, APCDs do not include all private payers. Consequently, it is not possible to distinguish a person who loses insurance coverage from a person who transitions to an insurance plan not captured in the APCD. However, many insurance transitions are possible to ascertain provided plans submit claims to

the APCD, which makes APCDs a unique, albeit incomplete, data to study cancer care across multiple payers.

Our results have similarities and differences to Utah's¹⁵ (UT) linkage. We found similar linkage rates for uninsured patients at the time of diagnosis (68% UT; 75% CO), private insurance (92% UT; 89% CO), and Medicaid (97% UT; 99% CO). However, our linkage rate is much higher for Medicare (79% UT; 99% CO), underscoring the variability between state APCDs and registries and the importance of including multiple years in a linkage. It is possible that this reflects an error in the way insurance was coded in the registry, which deserves more research.²⁶ Contrary to the Utah linkage, our linkage rates were consistent across rural/urban populations and racial and ethnic groups.

A few limitations are noteworthy. Our study is confined to a single state, which limits the ability to make broader comparisons across states when evaluating the quality of the data. More research is needed to validate APCD-registry linkages across states as regulations for claims submission is not uniform. Finally, uninsured patients are perhaps the most likely patients to experience unfavorable cancer outcomes, but they are not included in APCDs. The only source of claims or more detailed clinical information on uninsured individuals to date is data obtained from health care providers, such as electronic health records. Because of their limitations, APCD data cannot be used to distinguish an uninsured individual from one who transitions to a plan that does not submit claims to the APCD.

An important next step in advancing cancer research is to create a data infrastructure that combines APCD data linked to central registries in multiple states using the same linkage methodology for comparison and validation purposes. We hope that more states will develop standardized APCDs to facilitate comparisons and more studies will evaluate data quality. These data will be instrumental in evaluating the effects of private coverage and plan generosity, coverage disruptions, differences in insurance coverage and reimbursement, and state policies on cancer outcomes.¹⁴

ACKNOWLEDGMENTS

We thank John Arend, Manager of the Colorado Central Cancer Registry, for conducting the linkage and Julia Entwistle for editorial assistance. This study was supported by a grant from the National Cancer Institute (NCI), R01CA22599 (Bradley and Perrailon, Principal Investigators). Perrailon, Bradley, and Liang were also supported by NCI P30CA046934. All analyses were conducted by the Population Health Shared Resource at the University of Colorado Comprehensive Cancer Center (P30CA046934).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

ORCID

Marcelo C. Perrailon  <https://orcid.org/0000-0003-3355-3720>

Rifei Liang  <https://orcid.org/0000-0003-3260-7213>

Lindsay M. Sabik  <https://orcid.org/0000-0002-0613-4929>

Cathy J. Bradley  <https://orcid.org/0000-0001-9429-6567>

REFERENCES

- National Cancer Institute. SEER-Medicare Linked Database; 2020. <https://healthcaredelivery.cancer.gov/seermedicare/>
- Noone A-M, Lund JL, Mariotto A, et al. Comparison of SEER treatment data with Medicare claims. *Med Care*. 2016;54(9):e55-e64. doi:10.1097/MLR.000000000000073
- Setoguchi S, Earle CC, Glynn R, et al. Comparison of prospective and retrospective indicators of the quality of end-of-life cancer care. *J Clin Oncol*. 2008;26(35):5671-5678. doi:10.1200/JCO.2008.16.3956
- Holmes JA, Carpenter WR, Wu Y, et al. Impact of distance to a urologist on early diagnosis of prostate cancer among black and White patients. *J Urol*. 2012;187(3):883-888. doi:10.1016/j.juro.2011.10.156
- Tsui J, DeLia D, Stroup AM, et al. Association of Medicaid enrollee characteristics and primary care utilization with cancer outcomes for the period spanning medicaid expansion in New Jersey. *Cancer*. 2019;125(8):1330-1340. doi:10.1002/cncr.31824
- Bradley CJ, Given CW, Roberts C. Race, socioeconomic status, and breast cancer treatment and survival. *J Natl Cancer Inst*. 2002;94(7):490-496. doi:10.1093/jnci/94.7.490
- McDermott CL, Fedorenko C, Kreizenbeck K, et al. End-of-life services among patients with cancer: evidence from cancer registry records linked with commercial health insurance claims. *J Oncol Pract*. 2017;13(11):e889-e899. doi:10.1200/JOP.2017.021683
- Gelber RP, McCarthy EP, Davis JW, Seto TB. Ethnic disparities in breast cancer management among Asian Americans and Pacific islanders. *Ann Surg Oncol*. 2006;13(7):977-984. doi:10.1245/ASO.2006.08.036
- Parikh-Patel A, White RH, Allen M, Cress R. Risk of cancer among rheumatoid arthritis patients in California. *Cancer Causes Control*. 2009;20(6):1001-1010. doi:10.1007/s10552-009-9298-y
- Penberthy L, McClish D, Pugh A, Smith W, Manning C, Retchin S. Using hospital discharge files to enhance cancer surveillance. *Am J Epidemiol*. 2003;158(1):27-34. doi:10.1093/aje/kwg108
- McCarthy D. State all-payer claims databases: tools for improving health care value, part 1 how states establish an APCD and make it functional. The Commonwealth Fund; 2020. <https://www.commonwealthfund.org/publications/fund-reports/2020/dec/state-apcds-part-1-establish-make-functional>
- Agency for Healthcare Research and Quality. All-payer claims databases. 18 States have, interest in developing an APCD; 2018. <https://www.ahrq.gov/data/apcd/index.html#:~:text=To%20date%2C>
- All Payer Claims Database Council. Interactive State Report Map; 2021. <https://www.apcdouncil.org/state/map>
- Yabroff KR, Reeder-Hayes K, Zhao J, et al. Health insurance coverage disruptions and cancer care and outcomes: systematic review of published research. *J Natl Cancer Inst*. 2020;112(7):671-687. doi:10.1093/jnci/djaa048
- Garvin JH, Herget KA, Hashibe M, et al. Linkage between Utah all payers claims database and central cancer registry. *Health Serv Res*. 2019;54(3):707-713. doi:10.1111/1475-6773.13114
- McCarthy D. State all-payer claims databases tools for improving health care value Part 2: The uses and benefits of state APCDs; 2020. <https://www.commonwealthfund.org/publications/fund-reports/2020/dec/state-apcds-part-2-uses-benefits>
- National Cancer Institute. Geographically underserved areas; 2020. <https://cancercontrol.cancer.gov/research-emphasis/health-disparities/underserved-areas>
- United States Census Bureau. 2019 American community survey statistics for income, poverty and health insurance available for States and Local Areas; 2021. <https://www.census.gov/newsroom/press-releases/2020/acs-1year.html>
- United States Department of Agriculture Economic Research Service. Rural-Urban Commuting Area Codes; 2020. <https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/>
- Health Resources & Services Administration. What is shortage designation?
- United States Department of Agriculture Economic Research Service. Overview; 2019. <https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/#map>
- United States Department of Agriculture Economic Research Service. Rural poverty & well-being; 2020. <https://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being/>
- Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc*. 1969;64(328):1183-1210. doi:10.1080/01621459.1969.10501049
- Herzog TN, Scheuren FJ, Winkler WE. *Data Quality and Record Linkage Techniques*. Springer Science & Business; 2007.
- National Cancer Institute. Download Match*Pro Software; 2021. <https://surveillance.cancer.gov/matchpro/download>
- Sabik LM, Bradley CJ. Understanding the limitations of cancer registry insurance data-implications for policy. *JAMA Oncol*. 2018;4(10):1432-1433. doi:10.1001/jamaoncol.2018.2436

How to cite this article: Perrailon MC, Liang R, Sabik LM, Lindrooth RC, Bradley CJ. The role of all-payer claims databases to expand central cancer registries: Experience from Colorado. *Health Serv Res*. 2022;57(3):703-711. doi:10.1111/1475-6773.13901