HSR Health Services Research

DOI: 10.1111/j.1475-6773.2010.01221.x
RESEARCH ARTICLE

Building Capacity to Assess Cancer Care in the Medicaid Population in New York State

Francis P. Boscoe, Deborah Schrag, Kun Chen, Patrick J. Roohan, and Maria J. Schymura

Objective. To link data from a central cancer registry with Medicaid enrollment and claims files in order to assess cancer care in an economically disadvantaged population.
Data Sources. Over 500,000 cancer patients diagnosed between 2002 and 2006 reported to the New York State Cancer Registry were linked with New York State Medicaid enrollment and claims records.
Study Design. A probabilistic linkage was performed between the two data sources. The resulting Medicaid and non-Medicaid populations were compared in terms of demographics and stage at diagnosis.
Data Collection Methods. Existing databases were used.
Principal Findings. One-quarter of cancer patients were enrolled in Medicaid at or near the time of cancer diagnosis. The Medicaid cohort was younger, more likely to be an ethnic minority, foreign born, never married, live in either an inner-city or remote rural area, and have a higher stage at diagnosis.
Conclusions. The linked dataset will permit detailed analysis of cancer treatment and cancer treatment disparities among historically understudied groups. The linkage has also resulted in improvements in Cancer Registry quality through the identification of errors and missing values. The linkage did present technical challenges in the form of immense file sizes not easily adaptable to desktop computers.

Key Words. Data linkage, cancer registration, Medicaid, economically disadvantaged populations

This paper describes a linkage between the New York State Cancer Registry (NYSCR) and the New York State (NYS) Medicaid program, both housed within the New York State Department of Health (NYSDOH). Medicaid, the state-based health program for individuals and families with low incomes and resources in the United States, insures approximately one-sixth of New York adults aged 18–64 and one-fifth of elderly over 65, most of whom are dually enrolled in Medicare and Medicaid. Deficiencies in cancer care are known to

disproportionately affect the poor as well as racial and ethnic minorities who are overrepresented in Medicaid (Agency for Healthcare Research and Quality 2005; Institute of Medicine 2005; Landon et al. 2007).

The Surveillance, Epidemiology and End Results (SEER)–Medicare linkage is the prototype for our effort (Warren et al. 2002; National Cancer Institute, Division of Cancer Control and Population Sciences, Health Services and Economics Branch 2009). This linkage has enabled the development of a resource containing information on over 3 million elderly persons with cancer, which has yielded a large number of influential publications describing the patterns and outcomes of cancer care in the United States (Warren et al. 2002; Hershman et al. 2007; Wong et al. 2007; Gooden et al. 2008; Morris et al. 2008; White et al. 2008). The SEER program collects demographic and diagnostic information for persons diagnosed with cancer, historically representing about 14 percent of the U.S. population and more recently expanded to cover 26 percent. SEER data provide detailed characterizations of cancer diagnosis, tumor stage, and initial treatment but do not track care longitudinally other than vital status. Linkage with Medicare claims for enrollees in fee-for-service plans enables characterization of covered health services longitudinally. A comparable match to Medicaid records has seldom been attempted owing to technical and procedural obstacles (Bradley, Given, and Roberts 2001; Bradley et al. 2007). These obstacles include the fact that each state's Medicaid program is distinct, that Medicaid enrollment can be discontinuous, and that managed care penetration might be high and thus limit the utility of the claims for ascertaining care. The few attempts to construct registry—Medicaid linkages at the state level include those in Michigan (Bradley, Given, and Roberts 2001, 2002, 2003), California (Perkins et al. 2001; Chan et al. 2006), Ohio (Koroukian et al. 2006a, 2006b), Washington (Ramsey et al. 2008), and Louisiana (Whitaker et al. 2009).

The NYSCR–Medicaid linkage has two primary objectives. The first, discussed in this paper, is to create a deidentified analytic dataset for use in ongoing research projects with extramural partners. We will be using this dataset, for example, to compare community-dwelling cancer patients by

Address correspondence to Francis P. Boscoe, Ph.D., New York State Cancer Registry, 150 Broadway, Suite 361, Menands, NY 12204; e-mail: fpb01@health.state.ny.us. Deborah Schrag, M.D., is with the Gastrointestinal Cancer Center, Dana Farber Cancer Institute, Boston, MA. Kun Chen, Ph.D., is with the Center for Outcomes and Policy Research, Dana Farber Cancer Institute, Boston, MA. Patrick J. Roohan, M.S., is with the New York State Department of Health, Office of Health Insurance Programs, Albany, NY. Maria Schymura, Ph.D., is with New York State Cancer Registry, Menands, NY.

Medicaid status, stage distribution, and receipt of quality care to identify whether racial and ethnic disparities exist within the Medicaid program. Findings will be related to timing and duration of enrollment, as those who enrolled in Medicaid in response to a cancer diagnosis might have quite different diagnostic or demographic characteristics than long-term enrollees.

The second objective is to assess the degree to which Medicaid claims data add value to the cancer treatment information already collected by the NYSCR, with an emphasis on the treatment of breast and colorectal cancer. The collection of detailed treatment information is a relatively new undertaking for most U.S. cancer registries that are not part of the SEER program; treatment data for New York are considered complete only for cases diagnosed beginning in 2003. Even for this "complete" data, the amount of missing information is typically more than double that seen in SEER. For example, 3.2 percent of colorectal cancer cases diagnosed in NYS between 2004 and 2006 have unknown surgery information, compared with 1.3 percent in SEER. In addition, there is potentially useful information in claims records that cancer registries do not collect at all, such as screening utilization, prescription medications, and nursing home services. Claims data also provide a means for independently verifying that information being collected by central registries is correct.

In this paper, we detail the data linkage process and the resulting analytic dataset. We report the percentage of cancer cases matching to the Medicaid enrollment files by cancer site, stage, age, marital status, race/ethnicity, and geography, and show how these differ in important ways from the cancer population generally. We then discuss ways in which the linkage enhanced the quality of registry data, some of which were unanticipated. Finally, we discuss file size issues as the single largest technical hurdle encountered.

Data and Methods

The NYSCR is the nation's second oldest central cancer registry in the United States and receives reports on approximately 100,000 newly diagnosed tumors each year, augmented with mortality information from New York City and NYS vital records and the National Death Index. It has achieved the highest level (gold) certification from the North American Association of Central Cancer Registries for data completeness, timeliness, and accuracy for each of the past 9 years. NYSCR data used in this project consisted of all reportable tumors, including those with benign and uncertain behaviors,

diagnosed between 2002 and 2006 among adult NYS residents ($n = 545{,}250$ tumors among 517,394 individuals). For the 5 percent of cases with more than one Social Security number (SSN) in the NYSCR, each SSN was treated as a separate record, increasing the likelihood of a match. The SSN associated with the most definitive source—usually the diagnosing hospital—was designated the primary SSN, while others were designated as alternate SSNs.

The NYS Medicaid program provides medical coverage to approximately 5 million New Yorkers of all ages unable to afford health care. Eligibility is governed by income, household size, and/or the presence of a disability. Medicaid data used in this linkage consisted of 6.5 million NYS enrollees between 2001 and 2008 who were 18 years old by 2008. Data were drawn from a data warehouse containing information on enrollment, eligibility, claims, and encounters information. Encounters are claim-like transactions collected for managed care enrollees.

Matching 5 years of Cancer Registry data to 8 years of Medicaid enrollment data ensured that at least 1 year of claims antecedent to cancer diagnosis and 2 years of claims subsequent to diagnosis were potentially available for each patient. This will facilitate comparisons between persons with stable and unstable Medicaid enrollment and those not enrolled in Medicaid.

Following a formal application process and a series of technical planning meetings with the NYS Medicaid program, the NYSCR received an encrypted file of the 6.5 million eligible persons via secure file transfer protocol (ftp). This was matched to the cancer data by NYSCR staff using a six-pass probabilistic match using IBM *QualityStage* software, version 7.0 (Alur et al. 2008). This program and its predecessors have a long history of use by the New York State Health Department for data linkage activities. The match conducted for this project, however, could have been accomplished with any number of low-cost or no-cost solutions, such as the Centers for Disease Control's *LinkPlus* software (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm).

A probabilistic match compares multiple field values with each assigned a weight based on the likelihood that a match in that field was not due to chance. The sum of the weights for the matched fields is used to define matches, nonmatches, and those requiring clerical review. Each pass compares a different subset of fields, and passes may also specify blocking variables for which an exact match is required. This matching approach was developed internally within the NYSCR for routine case processing and has been applied to numerous research projects. Conservative matching thresholds were applied in order to minimize false positives. That is, we considered it more important to make sure that every case in the matched database was indeed enrolled in

Table 1: Probabilistic Matching Scheme Used in the Registry–Medicaid Match

Pass	*Blocking Variables*	*Matching Variables*	*Clerical Review Required*	*% of Total Matches*	*Comment*
1	SSN Surname	First name Middle initial Birth date Sex	None	83	Primarily identifies exact matches
2	SSN	First name Middle initial Birth date Sex	Negligible	5	Primarily identifies women who have changed surnames through marriage
3	Date of birth Surname soundex*	Surname First name soundex SSN Sex	Moderate	6	Primarily identifies minor typos and spelling variations
4	SSN	Surname First name Birth date First initial to middle initial Middle initial to first initial Sex	Moderate	$\ll 1$	Primarily identifies those who use first and middle names interchangeably, or who go by their middle name
5	First name Last name	Date of birth SSN Sex	None	3	Primarily identifies otherwise strong matches with one or both records missing SSN
6	Date of birth Surname soundex	Surname First name SSN Sex	Significant	3	A small number of valid matches with mostly false positives. Similar to pass 3 but with a much lower matching threshold

*Soundex is a function that groups similar-sounding names based on their distinctive consonants.

Medicaid, even at the risk of failing to identify a few of the Medicaid-enrolled cases as such. The six passes are described in Table 1. Clerical review of questionable matches was performed by the first author. Most of the matches determined to be false positives were likely spouses, relatives, or twins.

Once matches were determined, the matched records were sent back to the Medicaid program to obtain detailed claims information and enrollment histories for these cases. An analytic dataset was then created consisting of detailed demographic, clinical, and treatment information for each of the 517,394 cancer patients, along with their Medicaid enrollment and claims history, if applicable.

RESULTS

Approximately 25 percent of the 517,394 cancer patients were found to match to at least one Medicaid registrant ID during the 2001–2008 period. Over 78 million claims were associated with these cases, with a median of 300 claims per case and an average of 600.

Only 3 percent of the records containing alternate SSNs matched, suggesting that the large majority of the alternate SSNs were invalid. Ten percent of the cases matched to more than one unique Medicaid ID. These mainly describe situations where an individual moved to a different county (most often, between New York City and a location outside New York City) and reenrolled in Medicaid, even though doing so is unnecessary to maintain coverage. The matches also revealed approximately 700 cases where gender did not agree. Based on a review of a large sample of these cases, a slight majority of the errors were in the NYSCR data; these records were corrected and retained in the analytic dataset.

The ultimate analytic dataset included 80 demographic, diagnostic and treatment fields. About 5 percent of the cancer patients had multiple primary tumors during the study period, yielding a total of 545,250 records. Project-specific case and tumor identification numbers were created, with the case identification number serving as the key variable used for linking to the detailed Medicaid claims and enrollment data.

The entire process described in this paper, from initial application for the Medicaid data to the production of the analytic dataset, required about 6 months.

The analytic dataset was used to assess the demographic and stage characteristics of Medicaid enrollees with cancer. Table 2 lists the percentage of adult (aged 18–64) and elderly (aged 65+) cancer patients who were enrolled in Medicaid, ranked by cancer site, for the most common cancer sites and site groupings that collectively account for over 92 percent of all tumors in the file. Cancers that are more common among persons of lower socioeconomic status, including those that are HPV, smoking, and alcohol related (Menvielle et al. 2007; Benard et al. 2008; Mouw et al. 2008; Clegg et al. 2009), were among those with the highest percentages enrolled in Medicaid. (Here and in all subsequent tables, the numbers of cases are sufficiently large to provide stable proportions.)

Table 3 compares the stage distribution between those enrolled and not enrolled in Medicaid for three common cancer sites. Medicaid enrollees have a higher stage distribution for each site, with breast cancer showing the most pronounced difference. Medicaid enrollees are also more likely to have tumors reported as unstaged to the Cancer Registry.

Table 2: Percentage of Cancer Patients Enrolled in Medicaid,* Adults and Elderly, in Descending Order by Cancer Site, New York State, 2002–2006

	% Enrolled	
Site	*Adults (18–64)*	*Elderly (65+)*
All sites combined†	**27.3**	**23.2**
Liver	52.2	33.6
Cervix	50.8	48.9
Anal	44.4	31.7
Larynx	42.9	27.9
Stomach	41.7	33.0
Esophagus	38.7	25.4
Oral	38.5	27.0
Hodgkin lymphoma	37.0	26.5
Lung	35.5	22.8
Multiple myeloma	35.5	28.8
Brain	32.8	29.8
Non-Hodgkin lymphoma	32.0	18.8
Leukemia	31.8	20.8
Pancreas	30.5	22.4
Rectum	30.0	27.7
Colon	29.7	27.0
Ovary	27.0	24.5
Kidney	26.5	21.4
Testis	25.8	25.4
Uterus	23.4	27.2
Female breast	22.2	24.4
Thyroid	22.1	23.6
Bladder	19.6	18.5
Prostate	17.0	18.5
Melanoma	8.6	9.9

*"Enrolled in Medicaid" means ever enrolled in Medicaid during the 2001–2008 period; it does not mean enrolled at the time of diagnosis.

†Sites with values that are higher than for all sites combined are overrepresented in the Medicaid population; sites with values lower than this figure are underrepresented.

The percentage of cancer cases enrolled in Medicaid also displays substantial differences by gender, race, ethnicity, and age (Table 4). Elderly women are more likely to be enrolled, but the gender difference among adults is negligible. Hispanics, blacks, and Asians with cancer are much more likely to be enrolled in Medicaid than whites, at proportions two to three times higher. A majority of Hispanic and Asian cancer patients and just under half of black cancer patients are in Medicaid. Among the six most common Asian subgroups, Japanese and Filipinos have a substantially lower percentage of Medicaid cases than Chinese, Koreans, Vietnamese, or Asian Indians, with

Table 3: Cancer Stage Distribution by Medicaid Enrollment Status,* New York State, 2002–2006

	Lung		*Colorectal*		*Female Breast*	
Stage†	*Medicaid*	*Non-Medicaid*	*Medicaid*	*Non-Medicaid*	*Medicaid*	*Non-Medicaid*
In situ	—	—	7.3	8.8	15.4	23.0
Local	15.3	18.7	27.7	32.8	42.0	47.0
Regional	22.0	22.7	35.4	33.4	27.5	20.5
Distant	45.3	42.8	16.5	14.8	6.6	3.3
Unknown	17.3	15.7	13.1	10.2	8.4	6.2

*Enrolled in Medicaid at any time during the 2001–2008 period.
†SEER Summary Stage 2000.

the value for Japanese just slightly above that for white non-Hispanics. Major Hispanic subgroups also display substantial variation in this measure, ranging from 48 percent enrolled for Cubans to 82 percent enrolled for Dominicans.

Table 4: Percentage of Cancer Patients Enrolled in Medicaid,* by Gender, Race/Ethnicity,† and Age, New York State, 2002–2006

	All Ages	*Adults (18–64)*	*Elderly (65+)*
All cases combined‡	25.1	27.3	23.2
Males	22.4	27.2	19.0
Females	27.5	27.4	27.6
White	16.8	16.9	16.7
Black	48.1	50.9	44.7
Asian or Pacific Islander	57.0	51.4	65.0
Chinese	65.2	61.4	69.6
Japanese	19.6	19.3	20.1
Filipino	37.6	25.7	57.0
Korean	57.2	48.1	69.0
Asian Indian	55.0	51.6	62.4
Vietnamese	62.2	58.5	67.5
Other/unknown	14.5	13.5	16.1
Hispanic	60.6	60.3	61.0
Mexican	51.4	56.0	39.0
Puerto Rican	65.1	66.8	63.5
Cuban	47.9	50.1	46.8
South/Central American	58.9	57.2	61.4
Dominican	82.3	81.0	84.2

*"Enrolled in Medicaid" means ever enrolled in Medicaid during the 2001–2008 period; it does not mean enrolled at the time of diagnosis.
†Race/ethnicity categories used here are mutually exclusive.
‡Race/ethnic groups with values that are higher than for all cases combined are overrepresented in the Medicaid population; sites with values lower than this figure are underrepresented.

Further stratifying these results into adults and elderly patients reveals additional distinctions. Overall, a greater share of adult cancer patients is enrolled in Medicaid than elderly patients, but among Asians the reverse is true. The differences are especially pronounced among Filipinos and are also large for Koreans, Asian Indians, and Vietnamese. These findings appear to reflect generational differences in socioeconomic status for these groups. At the opposite extreme are Mexicans, with younger Mexicans having much higher Medicaid enrollment than elderly Mexicans.

There are also substantial geographic differences in Medicaid enrollment among cancer patients (supporting information Figure 1). After restricting the data to white non-Hispanics in order to eliminate the confounding effects of race and ethnicity, the percentage enrolled in Medicaid ranges from 8 percent in Nassau County, a suburban county of New York City, to 38 percent in Brooklyn. Of the four highest-enrolled counties, two are in New York City (Brooklyn, Bronx) and two are rural upstate counties with high poverty rates. The five lowest-enrolled counties are all wealthy suburban counties of New York City.

Marital status and nativity are also related to Medicaid enrollment (data not shown). Forty-six percent of never-married cancer patients are enrolled in Medicaid, while only 15 percent of married patients are. For divorced/separated persons the value is 39 percent, and for widowed 29 percent. Half of foreign-born cancer patients are enrolled in Medicaid versus 20 percent of those born in the United States. This difference is driven almost entirely by non-Hispanic whites and Asians/Pacific Islanders; for blacks and Hispanics, Medicaid enrollment is similarly high regardless of place of birth.

For several site and race/ethnicity combinations, over 70 percent of cancer patients are enrolled in Medicaid. For Asians/Pacific Islanders, these sites are larynx (75.6 percent, the highest such combination), esophagus, and stomach. For Hispanics, the sites are anus, cervix, larynx, liver, and oral; for blacks, cervix (data not shown).

DISCUSSION

The linkage between the NYSCR and New York State Medicaid enrollment files yields a large dataset that is weighted toward persons who are young, foreign born, never married, a member of an ethnic or racial minority, a resident of the inner-city or rural fringe, and diagnosed with a site of cancer associated with lower socioeconomic status. For some combinations of these variables, a majority of cancer patients are enrolled in Medicaid—for liver and

cervical cancer among adults, for all cancers among adult and elderly Asians and Hispanics, and for all cancers among black adults. These strata have been comparatively neglected in the many studies using SEER–Medicare data.

Our findings also suggest that Medicaid enrollees with breast, colorectal, or lung cancer are, on the whole, diagnosed at a more advanced stage of disease than their non-Medicaid counterparts. These findings must be interpreted with caution because they do not account for the tendency for some patients to enroll in Medicaid as a consequence of cancer diagnosis, whether as a result of having no previous insurance or a result of losing private insurance due to unemployment. The staging differences between Medicaid and non-Medicaid patients are expected to attenuate once the data are classified into those enrolled before diagnosis (prediagnosis cases) and those enrolled near the time of diagnosis (peridiagnosis cases). Specifically, we expect that peridiagnosis cases will have a more advanced stage distribution than prediagnosis cases. This analysis remains as future work.

Our probabilistic matching approach incorporating SSN, first name, middle name, last names, date of birth, and gender is similar to the SEER–Medicare matching approach (Potosky et al. 1993) and yields considerably more matches than the more common alternative of using SSN alone. Our use of different combinations of blocking and matching variables over six passes is effective at locating common errors, including typographical errors in a single field, different surnames owing to change in marital status, the swapping of first and middle name, and the use of a spousal SSN (Table 1). Approximately 11 percent of the matches did not involve an exact SSN match; the large majority of these were cases where SSN was missing from one or both records. A small number of matches also arose from minor differences in SSNs resulting from typographical errors. These 11 percent are slightly skewed toward being female (about 5 percent more likely than the rest of the dataset) because women are more likely to be missing an SSN. The choice of a probabilistic approach over a deterministic approach most likely did not make a large difference, as researchers have consistently found that matching results are insensitive to this choice, as long as both are equally well conceived and thorough (Gomatam et al. 2002; Bradley et al. 2007).

The matching process informed the completeness and quality of several Cancer Registry data items and generated some unanticipated findings. The primary SSN, typically the one associated with the "best" reporting source (usually the diagnosing hospital), was roughly eight times more likely to yield a match than an alternate SSN. Still, there were enough alternate SSN matches to justify their inclusion in the match. These alternate SSNs were subsequently

reclassified as the primary SSN in the Registry. There were 707 cases in which gender did not agree between the two sources. A clerical review of these cases determined that the NYSCR was more likely to have the incorrect gender by about a 3-to-2 margin. In correcting these errors, we noted that a disproportionate number of the sampled cases were women classified as males with breast cancer. This prompted a separate review of all male breast cancer cases; over 10 percent were found to have the sex miscoded, meaning the previously reported male breast cancer rate was >10 percent too high. (Because the female breast cancer rate is >100 times higher than the male rate, the impact on that rate was negligible, <0.1 percent). The phenomenon by which random misclassification errors accrue to rarer categories has been gaining attention in the cancer registry community (Boscoe et al. 2009).

The match also identified four duplicate cases in the Cancer Registry, wherein a single Medicaid record matched to multiple Cancer Registry records. While this is an inconsequential number in terms of data analysis, the identification and resolution of duplicate records is a task taken very seriously by the NYSCR, as it influences data certification. Any process that is helpful in identifying overlooked duplicate records is therefore of value. In the other direction, 10 percent of the cancer cases, or roughly 15,000, matched to more than one unique Medicaid ID. This is a consequence of people reregistering for the program, usually in different counties but sometimes merely under different names and addresses. The number of such cases was more than Medicaid program staff had anticipated.

A preliminary review of the linked data also revealed instances where treatment information not reported to the NYSCR was captured in a Medicaid record. Such omissions increased with increasing time after diagnosis, suggesting a problem with insufficient update records. The first course of treatment does not follow a fixed time limit, and delays can occur for any number of reasons. For example, one elderly patient was not able to receive recommended rectal surgery for over a year after diagnosis because of an intervening heart attack, as revealed in the Medicaid data. As a result of this finding, a reminder was sent to all facilities on the importance of update records. In a future activity, NYSCR staff will contact individual facilities to verify omitted treatment information identified by Medicaid.

We note that while federal rules on the use of Medicaid data do not support their direct inclusion into disease registries, this is not the case in our project. Medicaid data were either used to clarify among contradictory data items already stored in the registry (as with SSN) or to flag an item for independent verification (as with gender or treatment).

The large file sizes resulting from the data linkage posed some technical challenges. As mentioned previously, the claims file associated with the 131,009 matched patients contains over 78 million records (69 million after removing matches more than 2 years before cancer diagnosis). In plain text format this file is 13 GB in size; reading the file into statistical analysis software such as SAS doubles its size. Querying the file has proven time consuming, and it has necessitated some testing of different querying approaches to see which was most efficient. For example, on a free-standing, nonnetworked Dell Optiplex 760 computer with dual 3.16 GHz processors, 3.3 GB of RAM, 250 GB free drive space, and no other applications or processes running, an optimally written query to return the claims records associated with breast cancer cases among black women over age 50 takes 21 minutes to execute. Given the number of ways we wish to explore this data, our practice has been to develop complex queries and allow them to run overnight. This approach obviously slows the analytic process. We raise these points because they would be expected to be encountered by any other large state undertaking a similar linkage.

The data linkage and subsequent analysis reported here were conducted entirely within the NYSDOH, and no substantial bureaucratic or institutional hurdles were encountered in the process. Indeed, both the NYSCR and NYS Medicaid Program actively seek additional research uses of their data. Whether this could easily be repeated in another state would depend on the research cultures and organizational structures of the corresponding institutions. While the preparation of the analytic file was a complex task, to repeat this with additional years of data is now routine, as the processes and procedures and computer scripts are all in place.

In order to share the linked data with our extramural partners at the Dana Farber Cancer Institute, IRB approval was obtained from both NYSDOH and DFCI. To maximize data security, the shared data elements were limited to the minimum required for their planned analysis, and dates were masked by adding or subtracting a small multiple of 7 days to all date fields within a case record. The masking of dates was not a formal requirement of the IRB or institutional policy, but it was deemed conservative and prudent for this project. Files were encrypted using 256-bit encryption and transmitted via commercial courier on password-protected electronic media. Future work will explore the practicality of making these linked data available to a wider research audience.

Linking data from the NYSCR and the New York State Medicaid program has proven to be technically complex but manageable and rewarding, based on preliminary analyses. The linkage highlights population cohorts that

have been comparatively understudied and represents a valuable resource for future comparative effectiveness research. Medicaid patients have more advanced stage at diagnosis than their non-Medicaid counterparts, with future work to determine how much of this effect is related to the timing and duration of enrollment. Future work will also assess the extent to which Medicaid patients are receiving quality cancer care as measured by national guidelines, and whether racial and ethnic disparities in care exist. The linkage has already prompted a number of improvements in the quality of demographic data items, and in time this will be expanded to diagnostic and treatment data items.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported by Cooperative Agreement S3888 from the American Schools of Public Health/ Centers for Disease Control and Prevention and National Cancer Institute R01-CA131847-01A1.

Disclaimers: Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or NCI. The authors acknowledge Peter Gallagher, Nicholas Asimakopoulos, and Marlene Gordon for their assistance in data compilation and linkage, and Amber Sinclair, Gail Samuelson, Eric Tai, and Lisa Richardson for editorial review.

Disclosures: The authors claim no conflicts of interest.

REFERENCES

Agency for Healthcare Research and Quality. 2005. *2005 National Healthcare Disparities Report. AHRQ Publication No. 06-0017.* Rockville, MD: Agency for Healthcare Research and Quality.

Alur, N., A. K. Jha, B. Rosen, and T. Skov. 2008. *IBM WebSphere QualityStage Methodologies, Standardization, and Matching.* Poughkeepsie, NY: IBM Corp.

Benard, V. B., C. J. Johnson, T. D. Thompson, K. B. Roland, S. M. Lai, V. Cokkinides, F. Tangka, N. A. Hawkins, H. Lawson, and H. K. Weir. 2008. "Examining the Association between Socioeconomic Status and Potential Human Papillomavirus-Associated Cancers." *Cancer* 113: 2910–8.

Boscoe, F. P., M. J. Schymura, M.-C. Hsieh, M. A. Williams, and K. A. Henry. 2009. "Issues with the Coding of Pacific Islanders in Central Cancer Registries." *Journal of Registry Management* 35: 47–51.

Bradley, C. J., C. W. Given, and C. Roberts. 2001. "Disparities in Cancer Diagnosis and Survival." *Cancer* 91: 178–88.

———. 2002. "Race, Socioeconomic Status, and Breast Cancer Treatment and Survival." *Journal of National Cancer Institute* 94: 490–6.

———. 2003. "Late Stage Cancers in a Medicaid-Insured Population." *Medical Care* 41: 722–8.

Bradley, C. J., C. W. Given, Z. Luo, C. Roberts, G. Copeland, and B. A. Virnig. 2007. "Medicaid, Medicare, and the Michigan Tumor Registry: A Linkage Strategy." *Medicine and Decision Making* 27: 352–63.

Chan, J. K., S. L. Gomez, C. D. O'Malley, C. I. Perkins, and C. A. Clarke. 2006. "Validity of Cancer Registry Medicaid Status against Enrollment Files: Implications for Population-Based Studies of Cancer Outcomes." *Medical Care* 44: 952–5.

Clegg, L. X., M. E. Reichman, B. A. Miller, B. F. Hankey, G. K. Singh, Y. D. Lin, M. T. Goodman, C. F. Lynch, S. M. Schwartz, V. W. Chen, L. Bernstein, S. L. Gomez, J. J. Graff, C. C. Lin, N. J. Johnson, and B. K. Edwards. 2009. "Impact of Socioeconomic Status on Cancer Incidence and Stage at Diagnosis: Selected Findings from the Surveillance, Epidemiology, and End Results: National Longitudinal Mortality Study." *Cancer Causes Control* 20: 417–35.

Gomatam, S., R. Carter, M. Ariet, and G. Mitchell. 2002. "An Empirical Comparison of Record Linkage Procedures." *Statistics in Medicine* 21: 1485–96.

Gooden, K. M., D. L. Howard, W. R. Carpenter, A. P. Carson, Y. J. Taylor, S. Peacock, and P. A. Godley. 2008. "The Effect of Hospital and Surgeon Volume on Racial Differences in Recurrence-Free Survival after Radical Prostatectomy." *Medical Care* 46: 1170–6.

Hershman, D., A. I. Neugut, J. S. Jacobson, J. Wang, W. Y. Tsai, R. McBride, C. L. Bennett, and V. R. Grann. 2007. "Acute Myeloid Leukemia or Myelodysplastic Syndrome Following Use of Granulocyte Colony-Stimulating Factors during Breast Cancer Adjuvant Chemotherapy." *Journal of National Cancer Institute* 99: 196–205.

Institute of Medicine. 2005. *Assessing the Quality of Cancer Care: An Approach to Measurement in Georgia.* Washington, DC: National Academies Press.

Koroukian, S. M., H. Beaird, E. Madigan, and M. Diaz. 2006a. "End-of-Life Expenditures by Ohio MEDICAID Beneficiaries Dying of Cancer." *Health Care Financing Review* 28: 65–80.

Koroukian, S. M., F. Xu, A. Dor, and G. S. Cooper. 2006b. "Colorectal Cancer Screening in the Elderly Population: Disparities by Dual Medicare-Medicaid Enrollment Status." *Health Services Research* 41: 2136–54.

Landon, B. E., E. C. Schneider, S. L. Normand, S. H. Scholle, L. G. Pawlson, and A. M. Epstein. 2007. "Quality of Care in Medicaid Managed Care and Commercial Health Plans." *JAMA* 298: 1674–81.

Menvielle, G., A. E. Kunst, I. Stirbu, C. Borrell, M. Bopp, E. Regidor, B. H. Strand, P. Deboosere, O. Lundberg, A. Leclerc, G. Costa, J. F. Chastang, S. Esnaola, P. Martikainen, and J. P. Mackenbach. 2007. "Socioeconomic Inequalities in Alcohol Related Cancer Mortality among Men: To What Extent Do They Differ

between Western European Populations?" *International Journal of Cancer* 121: 649–55.

Morris, A. M., K. G. Billingsley, A. J. Hayanga, B. Matthews, L. M. Baldwin, and J. D. Birkmeyer. 2008. "Residual Treatment Disparities after Oncology Referral for Rectal Cancer." *Journal of National Cancer Institute* 100: 738–44.

Mouw, T., A. Koster, M. E. Wright, M. M. Blank, S. C. Moore, A. Hollenbeck, and A. Schatzkin. 2008. "Education and Risk of Cancer in a Large Cohort of Men and Women in the United States." *PLoS ONE* 3: e3639.

National Cancer Institute, Division of Cancer Control and Population Sciences, Health Services and Economics Branch. 2009. "SEER-Medicare Data" [accessed on August 18, 2009]. Available at http://healthservices.cancer.gov/seermedicare/overview/seermed_fact_sheet.pdf

Perkins, C. I., W. E. Wright, M. Allen, S. J. Samuels, and P. S. Romano. 2001. "Breast Cancer Stage at Diagnosis in Relation to Duration of Medicaid Enrollment." *Medical Care* 39: 1224–33.

Potosky, A. L., G. F. Riley, J. D. Lubitz, R. M. Mentnech, and L. G. Kessler. 1993. "Potential for Cancer-Related Health-Services Research Using a Linked Medicare-Tumor Registry Database." *Medical Care* 31: 732–48.

Ramsey, S. D., S. B. Zeliadt, L. C. Richardson, L. Pollack, H. Linden, D. K. Blough, and N. Anderson. 2008. "Disenrollment from Medicaid after Recent Cancer Diagnosis." *Medical Care* 46: 49–57.

Warren, J. L., C. N. Klabunde, D. Schrag, P. B. Bach, and G. F. Riley. 2002. "Overview of the SEER–Medicare Data: Content, Research Applications, and Generalizability to the United States Elderly Population." *Medical Care* 40: IV 3–18.

Whitaker, L. M., R. Young, K. Poche, G. Bucher, P. Soto, J. P. Martinez, M. S. Hendrix, X. C. Wu, L. Tichenor, and V. W. Chen. 2009. Improving First Course Treatment and Follow Up Information with Medicaid Linkage. Presented at *North American Association of Central Cancer Registries Annual Conference*, San Diego, CA, June 13–19.

White, A., C. C. Liu, R. Xia, K. Burau, J. Cormier, W. Y. Chan, and X. L. L. Du. 2008. "Racial Disparities and Treatment Trends in a Large Cohort of Elderly African Americans and Caucasians with Colorectal Cancer, 1991 to 2002." *Cancer* 113: 3400–9.

Wong, S. L., H. Ji, B. K. Hollenbeck, A. M. Morris, O. Baser, and J. D. Birkmeyer. 2007. "Hospital Lymph Node Examination Rates and Survival after Resection for Colon Cancer." *JAMA* 298: 2149–54.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Figure S1: Percentage of White Non-Hispanic Cancer Patients Enrolled in Medicaid, by County, New York State, 2002–2006.