Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world-wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____	_____
Kirsten M. Woolpert	4/22/2020

Validation of LexisNexis® Accurint® in the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program

By

Kirsten M. Woolpert
Master of Public Health

Department of Epidemiology

_____

Timothy L. Lash
Committee Chair

Validation of LexisNexis® Accurint® in the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program

By

Kirsten M. Woolpert

Bachelor of Science

University of North Carolina at Wilmington

2018

Faculty Thesis Advisor: Timothy L. Lash, DSc, MPH

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2020

Abstract

Validation of LexisNexis® Accurint® in the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program

By Kirsten M. Woolpert

Background: LexisNexis® Accurint® is a database of ~45 billion public records that provides information regarding individual's location of residence. This low-cost method can be used as a tool in prospective cohort studies to maintain high follow-up rates, but to date has not been validated to ensure accuracy in tracking individuals. This study utilized the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program (CRISP) of 69,494 cancer patients to validate the software and to examine predictors of patients not included or who had an inaccurate entry in LexisNexis.
Methods: Cancer patients within the Georgia Cancer Registry are routinely linked to the National Death Index (NDI), providing for decedents the US state in which the patient died. We compared the state of residence reported in Lexis Nexis with the NDI state of residence at death as the gold standard, allowing for calculations of sensitivity and specificity of state of residence information in Lexis Nexis. Additionally, multivariate logistic regression analyses were performed to examine associations between demographic information provided through the registry and three outcomes: 1. having a match between LexisNexis and NDI, 2. being missed in the LexisNexis database, and 3. moving out of the state of Georgia according to LexisNexis.
Results: Of the 69,494 patients in the CRISP cohort, 65,890 (94.8%) were found in LexisNexis, and a total of 9,597 (13.8%) had died. The sensitivity of the LexisNexis software for identifying persons who moved out of Georgia was 34.6% and the specificity was 89.3%. Unmarried individuals, blacks, Asian/Pacific Islanders, Hispanics, individuals living in high poverty neighborhoods, and younger patients were all more likely to be missed in the LexisNexis database as well as to have a discordance between LexisNexis state of residence and the National Death Index state of residence at death.
Discussion: This study showed that LexisNexis Accurint did not accurately identify state of residence at death in a large proportion of CRISP cohort members. Since achieving high follow-up rates is essential in any prospective cohort, the low validity of this software is important to note for researchers planning to use this software for follow-up. The generalizability of results to persons who had not died by end of follow-up is an important consideration.

Validation of LexisNexis® Accurint® in the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program

By

Kirsten M. Woolpert

Bachelor of Science

University of North Carolina at Wilmington

2018

Faculty Thesis Advisor: Timothy L. Lash, DSc, MPH

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2020

ACKNOWLEDGEMENTS

I would like to thank Dr. Timothy Lash, the faculty thesis advisor of this work, for all of his guidance and support in completing this study. Additionally, I would like to thank Dr. Kevin Ward and Cameron England for allowing me the opportunity to get involved in the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program, as well as for all their help in supporting this project.

TABLE OF CONTENTS

BACKGROUND

With an estimated 15.5 million cancer survivors living in the United States in 2016, cancer researchers are faced with a growing population with many unique health needs (1). A major gap in the field of cancer epidemiology is the lack of population-wide recurrence data. Although U.S. cancer registries collect information on a variety of important variables that can assess cancer burden and describe patterns of survival, a population-wide registry that reports cancer recurrence does not currently exist (2, 3). In 2019, the Georgia Cancer Registry (GCR) became a part of an NCI funded grant exploring ways to automate the capture of recurrence data. This methodology involves the use of data streams to build algorithms that generate signals of recurrence which will be validated by field staff. The Georgia cancer recurrence cohort will consist of patients with a first primary non-metastatic and invasive breast, prostate, lymphoma, or colorectal cancer diagnosis between 2013 and 2017, all of which will be followed for recurrence of their disease.

A concern with validity in this cohort is with the potential for loss to follow-up over the study period. Since the data streams utilized in this study are mostly exclusive to Georgia, properly censoring individuals who leave the state will be essential. Based on cancer mortality data in the GCR, approximately six percent of cancer deaths among Georgia patients occur in another state (4). These patients should be censored when they emigrate from Georgia, but there is no systematic strategy to identify their emigration. As the cohort begins collecting data on the outcome status of patients, there could be concerns with differential loss to follow-up and systematically missing data on recurrence status. In a prior European study by Ginsburg et al. on loss to follow-up in an active surveillance system of prostate cancer, it was found that not only were African American patients at a higher risk of having poorer prostate cancer outcomes, but

they were also more likely to be lost to follow-up (5). If we can identify patients at a high risk of being lost to follow-up in Georgia's cancer recurrence cohort, we can develop strategies to ensure proper censoring of their follow-up. Additionally, we can use the knowledge we learn about patient characteristics associated with not linking to the LexisNexis database to explore other possible follow-up opportunities.

To offset this concern and maintain high follow-up rates, cohort members will be annually linked to LexisNexis® Accurint®, a database of ~45 billion public records that provides information that includes individual's location of residence (6). This software uses information such as bankruptcy records, motor vehicle registrations, and personal property records to identify and track individuals. As a low-cost method, this tool may be a powerful means to track cohort members even after leaving the study's catchment area. Prior research has explored the utilization of this software for identifying control populations; however, no study to our knowledge has explored how well this software works as a tool for longitudinally following where individuals reside (7). The aim of this study is to determine the validity of LexisNexis Accurint in accurately tracking the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program (CRISP) cohort members. Can one use this tool to accurately identify when a Georgia resident leaves the state so they can be appropriately censored? Since the GCR does not currently know which patients may have moved elsewhere, this validation will be done using cohort members who have died, allowing for the use of the state at the time of death recorded in the National Death Index of the National Center for Health Statistics as the gold standard measurement. If an individual's residence at time of death was in another state, that individual had to leave the state of Georgia at some point in time. With this information, the sensitivity and specificity of LexisNexis residential data can be calculated, at least among those

who have died, allowing for potential quantitative bias analysis should LexisNexis fail to correctly classify all cohort members. In addition, this validation study will assess the predictors of cases whose residence state in LexisNexis does not match their state of death from the NDI, as well as those who were not included in the LexisNexis database at all.

METHODS

Study Population

The recurrence surveillance cohort currently consists of Georgia cancer patients with a first primary non-metastatic and invasive breast, prostate, lymphoma, or colorectal cancer diagnosis between 2013 and 2017. Over this five-year period, there were a total of 27,453 breast cancer patients, 12,067 colorectal cancer patients, 5,088 lymphoma patients, and 24,886 prostate cancer patients identified. This cohort of 69,494 Georgia cancer patients represents the first group of those who will be followed in the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program (CRISP) to determine cancer recurrence risks and rates. To address the possibility of patients in our cohort being lost to follow-up due to leaving the state of Georgia, the recurrence surveillance cohort will be annually linked to LexisNexis Accurint, a database of public records used to follow the state of residence of cohort members.

Validating LexisNexis® Accurint® against the National Death Index

The National Death Index is a centralized index of death record information routinely collected across the United States (8). Linkages of GCR data with death information from the NDI are routinely performed and the data from the NDI are integrated into the registry. For this analysis, the NDI was the gold standard measurement, providing the state of residence for each individual at their time of death.

The entire CRISP cohort was linked to LexisNexis, which provides the best-known and most current address for each individual per the batch configuration established by the Georgia Cancer Registry. An initial assumption was made that the most current address from LexisNexis should correspond with the address where the patient was residing at death per the NDI.

Comparisons could then be made between the state of residence at death provided by the National Death Index and the residential state provided by LexisNexis among the cohort members at the time of their death (Figure 1).

Manual Review in LexisNexis Accurint

Upon initial review of the data, it became apparent that the most current address provided by LexisNexis can be a different address (i.e. in a different state) from where the patient resided at the time of death. We assume, but cannot confirm with certainty, that LexisNexis must continue to track some individuals even after the death, likely by following addressed of family members and spouses, in order to estimate a best address for continued follow-up (e.g., to resolve outstanding financial obligations). Because this negated our initial assumption about the most current address provided by LexisNexis, a manual review of records was performed for selected individuals as described below. LexisNexis can be searched in a more time-consuming manual fashion which allows the user to see an address history for each individual with estimated dates at each given residence. Cases were manually examined and updated where the National Death Index recorded death in Georgia, but the most current address from LexisNexis was out of state. Additionally, cases were also studied where the National Death Index recorded death out of state, but the most current address from LexisNexis was in Georgia. Both reviews were conducted by accessing the cases' LexisNexis longitudinal residential address history and examining the individual's state of residence prior to the date of their death. It was assumed in these analyses that concordance between the most current residential state from LexisNexis and the residence state at time of death from NDI represented an accurate result for LexisNexis. These cases were not manually reviewed.

Statistical Analysis

Sensitivity was defined as the probability of being correctly classified as residing outside the state of Georgia in the LexisNexis database for those who died out of state. Specificity was defined as the probability of being correctly classified as residing within the state of Georgia in the LexisNexis database for those who died in state. Individuals who LexisNexis identified as being out of the state of Georgia at the time of death, but the NDI recorded a different out of state address were considered to have a mismatch and were included in the sensitivity denominator. Using demographic information provided in the Georgia Cancer Registry, we evaluated characteristics of the cohort including sex, marital status, Hispanic origin, race, and poverty level. Poverty level was estimated using the census tract poverty indicator variable, which is an indicator for neighborhood poverty level based on the census tract of the residential address at initial cancer diagnosis (9). Among the entire cohort, we examined these characteristics in association with two outcomes: 1. being missed entirely in the LexisNexis database, and 2. moving out of the state of Georgia according to LexisNexis. Among those who have died, we examined these characteristics in association with having a match between LexisNexis and the National Death Index. This analysis was done using multivariate logistic regression including each variable of interest. All models were adjusted for sex, marital status, race, Hispanic origin, census tract poverty indicator, age at diagnosis, and vital status. We conducted analyses using SAS version 9.4 (SAS Institute, Cary, NC).

RESULTS

Of the 69,494 patients in the CRISP cohort, 65,890 (94.8%) were linked with information in the LexisNexis database. According to LexisNexis, 4,588 (7.0%) of these 65,890 patients had a current out of state address. Among the 9,597 cohort members who had died, 8,278 (86.3%) had a matching state of death between LexisNexis and the National Death Index (Table 1). When exploring the discordant match status between LexisNexis and the NDI, there were a total of 795 cases that needed to be manually reviewed. After redefining the state of death of these 795 cases in LexisNexis to reflect the individual's most recent residence before death rather than the concurrent estimated best address, there were 471 (59.2%) cases that still had a mismatch when comparing the two sources of information. After the manual review, matches between LexisNexis and NDI increased, thus improving sensitivity and specificity (Table 2). Using the NDI as the gold standard after the manual review, the sensitivity of the LexisNexis software was 34.6% (n=161/466; 95% CI: 30.3%, 39.0%) and the specificity was 89.3% (n=8,111/9,086; 95% CI: 88.6%, 89.9%).

When exploring the predictors of having a match between NDI and LexisNexis, various demographic characteristics were examined (Table 1). Controlling for age at diagnosis, census tract poverty indicator, sex, race, and marital status, Hispanics had higher odds of having a mismatch when compared to Non-Hispanics (OR: 2.8, 95% CI: 2.1, 3.7). Additionally, unmarried individuals were more likely to not have a matching state between LexisNexis and NDI. Single patients had nearly three times the odds of having a mismatching state compared to married patients (OR: 2.8, 95% CI: 2.4, 3.3). Odds of a mismatching state were also higher among blacks and among Asian/Pacific Islanders when compared to whites, and slightly higher among males when compared to females.

In the entire CRISP cohort, there were 3,604 participants who were not found in the LexisNexis software (Table 3). Controlling for sex, marital status, race, census tract poverty indicator, Hispanic origin, age at diagnosis, and vital status, cohort members who had died were more likely to be missed by LexisNexis (OR: 2.1, 95% CI: 1.9, 2.3). Individuals of Hispanic origin had an increased odds of not being included in the LexisNexis database (OR: 5.5, 95% CI: 4.9, 6.1), and people who are single were more likely to be missed than those who are married (OR: 4.0, 95% CI: 3.6, 4.3). Males were more likely to not have any information in LexisNexis when compared to females, and patients aged 15-44 had higher odds of being missed compared to those aged 55-64. Compared to white patients, Blacks and Asian/Pacific Islanders also had an increased odds of being missed in LexisNexis. Finally, patients living in higher poverty neighborhoods according to the census tract of diagnosis address were more likely to be missing in LexisNexis compared to those living in neighborhoods with 0-<5% poverty.

There were 4,588 individuals in the CRISP cohort who left the state of Georgia, and 61,302 who stayed in Georgia according to LexisNexis data (Table 4). Individuals who were not married were more likely to leave the state, controlling for sex, race, Hispanic origin, census tract poverty indicator, age at diagnosis, and vital status. Controlling for these same variables, Hispanics were more likely than non-Hispanics to leave Georgia after diagnosis (OR: 1.8, 95% CI: 1.6, 2.1). Additionally, patients living in higher poverty neighborhoods were less likely to leave the state than those living in low poverty areas (OR: 0.6, 95% CI: 0.6, 0.7).

DISCUSSION

This study showed that while LexisNexis Accurint was able to identify a most recent residential address for a majority of CRISP cohort members, among those who had died, the sensitivity and specificity of the software was low. The software was able to accurately classify 34.6% of individuals dying outside of the state of Georgia, and 86.3% of individuals dying within the state of Georgia. In particular, the low sensitivity of the software demonstrated that LexisNexis was not able to identify the majority of patients who actually left the state of Georgia, at least among those who were deceased. Being that achieving high follow-up rates is essential to any cohort study, the low validity of LexisNexis is important to note for any researchers who use this software tool, at least among deceased individuals and possibly generalizing to all cohort members. In the context of the Georgia Cancer Registry's CRISP cohort, this study showed that some groups may be more likely to leave the state and be lost to follow-up. These groups included unmarried individuals, American Indians/Alaskan Natives, unknown race, Hispanics, people that lived in a low poverty area at initial cancer diagnosis, and cancer cases under 44. This is important to note as the study begins collecting information on recurrence status, as there could be a selection bias introduced by systematically missing these groups of people.

One of the caveats to using LexisNexis to follow cohort members who have died is that the software provides a best guess for an individual's address rather than their most recent location. In this study, a manual review was conducted to look further into this, where it was found that LexisNexis does not appear to routinely capture death data and that they may be following family members and spouses of the actual decedent, even after the date of death. While obtaining longitudinal data from LexisNexis is possible, the software provides a range of dates for each address, which may make it challenging to incorporate into long-term cohort

studies. This manual review did improve the sensitivity and specificity of the software, but still showed that there are limitations to using LexisNexis among cases who have died.

Another finding of this study was that LexisNexis had a higher probability of inaccurately classifying state of death among males, unmarried individuals, blacks and Asian/Pacific Islanders, people of Hispanic origin, people living in census tract neighborhoods with a higher poverty level, and cases who are diagnosed at a younger age. Additionally, these same groups with the addition of American Indians/Alaskan Natives, people of unknown race, cases diagnosed above the age of 75, and cases who have died were more likely to not be included in the LexisNexis system at all. These groups of people are conventionally the groups that researchers conducting cohort studies try to maintain high follow-up with, as they typically are more likely to be lost to follow-up as well as to experience adverse health outcomes (10, 11). The study then highlights the need for the development of continued methods to follow these patients, as LexisNexis did not appear to proportionately follow these groups.

Though LexisNexis Accurint can be a powerful tool for maintaining high follow-up rates in prospective cohort studies, this study demonstrated that its ability to accurately track state of residence at an individual's death may be limited. Because LexisNexis provides the best address and not always the most recent address before death, researchers should be cautious about using this tool for follow-up on deceased patients. Additionally, this software showed poorer follow-up for many minority groups, either by inaccurately classifying the state of their death or by having no information on them. However, this validation study likely does not represent the experience of following cohort members who are alive. Future studies could develop a methodology to explore the validation of using LexisNexis to follow the location of living individuals.

STRENGTHS AND WEAKNESSES

One of the major strengths in this study was the utilization of the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program. This dataset of 69,494 individuals not only provided a large cohort of cases to perform this research, but through the cancer registry, most demographic information was readily available allowing for very low rates of missing data. We were able to link the majority of this cohort to LexisNexis, which also provided a large sample size to conduct this validation study. Second, we had access to individual records in the LexisNexis system. When it was found that the latest residential address was not always indicative of where the cancer case had died, we used the LexisNexis software to perform a manual review, allowing for a more comprehensive understanding of how LexisNexis works and its validity.

This validation study did have some limitations. Ideally, we would have performed a manual review in LexisNexis of all 9,597 deceased cases. This would have allowed us to ensure that the state of residence at the time of an individual's death truly was the state where the individual was residing at the time of their death. However, our approach of validating the 795 cases that had a discordance between National Death Index and LexisNexis was most feasible and provided a best-case scenario for what the sensitivity and specificity of this software was. Since the software was still found to be inaccurate even after this review, it likely was not necessary to manually review all cases. Additionally, this study did not allow us to validate the LexisNexis software among the living cancer survivors. This validation study was using the National Death Index as the gold standard, which limited our analysis to cancer cases who had died. Many cancer patients relocate when their cancer progresses to seek treatment directed care, live nearer to family, or to receive palliative care. These relocation forces would not affect cohort

patients who remained cancer free. Though we would expect the software to more accurately track living individuals, we were not able to confirm this through this study.

FUTURE DIRECTIONS

Because the gold standard in this analysis was the National Death Index, this validation study was limited to cancer patients who died. While this was an important first step in studying the validity of utilizing this software in the cohort setting, we were unable to determine how well LexisNexis works among living individuals. Future studies could explore this by developing a method to compare the state of residence in LexisNexis to another reliable source that is not limited to deceased cases. Additionally, it may be of interest to validate this software in other populations, especially prospective cohort studies that are using LexisNexis as a method of follow-up.

REFERENCES

1. American Cancer Society. Cancer Facts & Figures 2019. Atlanta: American Cancer Society. 2019.
2. In H, Bilimoria KY, Stewart AK, et al. Cancer Recurrence: An Important but Missing Variable in National Cancer Registries. Annals of Surgical Oncology. 2014;21(5):1520-9. doi:10.1245/s10434-014-3516-x
3. Warren JL, Yabroff KR. Challenges and Opportunities in Measuring Cancer Recurrence in the United States. JNCI: Journal of the National Cancer Institute. 2015;107(8). doi:10.1093/jnci/djv134
4. McNamara C BA, Ward KC. Georgia Cancer Data Report 2016. Georgia Department of Public Health, Georgia Comprehensive Cancer Registry. 2016.
5. Ginsburg KB, Auffenberg GB, Qi J, et al. Risk of Becoming Lost to Follow-up During Active Surveillance for Prostate Cancer. European Urology. 2018;74(6):704-7. doi:10.1016/j.eururo.2018.08.010
6. Accurint®, a LexisNexis® commercial location and research online service. Philadelphia, PA: LexisNexis, 2005. (http://www.accurint.com/).
7. Stone MB, Lyon JL, Simonsen SE, White GL, Jr., Alder SC. An Internet-based Method of Selecting Control Populations for Epidemiologic Studies. American Journal of Epidemiology. 2006;165(1):109-12. doi:10.1093/aje/kwj351
8. Centers for Disease Control and Prevention. National Death Index User's Guide. Accessed from https://www.cdc.gov/nchs/data/ndi/NDI_Users_Guide.pdf.
9. North American Association of Central Cancer Registries. Version 16 Data Standards and Data Dictionary. http://datadictionary.naaccr.org.

10. Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to Follow-up in Cohort Studies: Bias in Estimates of Socioeconomic Inequalities. Epidemiology. 2013;24(1):1-9. doi:10.1097/EDE.0b013e31827623b1

11. Knudsen AK, Hotopf M, Skogen JC, Overland S, Mykletun A. The health status of nonparticipants in a population-based health study: the Hordaland Health Study. Am J Epidemiol. 2010;172(11):1306-14. doi:10.1093/aje/kwq257

TABLE 1. Match Status between Residential State at Time of Death in LexisNexis Accurint and State of Death in National Death Index (NDI) among Cancer Recurrence and Information Surveillance Program (CRISP) Members

Characteristic	**State Match between LexisNexis and NDI[a] (n=8,278)** n (%)	**No State Match between LexisNexis and NDI[b] (n=1,319)** n (%)	**Adjusted OR[c]** OR (95% CI)
Sex			
Male	3,842 (46.4)	639 (48.5)	1.1 (1.0, 1.2)
Female	4,436 (53.6)	680 (51.6)	Reference
Marital status[d]			
Single	1,197 (14.5)	370 (28.1)	2.8 (2.4, 3.3)
Married	3,766 (45.5)	382 (29.0)	Reference
Separated	97 (1.2)	24 (1.8)	2.2 (1.4, 3.5)
Divorced	876 (10.6)	144 (10.9)	1.7 (1.4, 2.1)
Widowed	1,635 (19.8)	274 (20.8)	2.1 (1.7, 2.5)
Unknown	695 (8.4)	123 (9.3)	1.8 (1.4, 2.2)
Race			
White	5,511 (66.6)	745 (56.5)	Reference
Black	2,661 (32.2)	519 (39.4)	1.4 (1.3, 1.6)
American Indian/Alaskan Native	<5	<5	-
Asian or Pacific Islander	99 (1.2)	49 (3.7)	3.6 (2.5, 5.1)
Unknown	<5	<5	-
Hispanic origin			
Non-Hispanic	8,092 (97.8)	1,242 (94.2)	Reference
Hispanic	186 (2.3)	77 (5.8)	2.8 (2.1, 3.7)
Census Tract Poverty Indicator[e]			
0%-<5% Poverty	718 (8.7)	100 (7.6)	Reference
5%-<10% Poverty	1,336 (16.1)	225 (17.1)	1.2 (0.9, 1.6)
10%-<20% Poverty	2,844 (34.4)	430 (32.6)	1.1 (0.8, 1.3)
20%-100% Poverty	3,380 (40.8)	564 (42.8)	1.1 (0.9, 1.4)
Age at Diagnosis			
15-44	401 (4.8)	114 (8.6)	1.3 (0.8, 1.9)
45-54	822 (9.9)	146 (11.1)	0.9 (0.7, 1.2)
55-64	1,564 (18.9)	279 (21.2)	Reference
65-74	2,478 (29.9)	358 (27.1)	0.9 (0.7, 1.1)
75+	3,013 (36.4)	422 (32.0)	1.0 (0.7, 1.4)

a. Includes individuals where both NDI and LexisNexis recorded state of death at time of death as GA, a matching out of state address, or unknown.
b. Includes individuals where NDI and LexisNexis recorded different states at time of death, or one program recorded state of death as unknown/missing.
c. Multivariate logistic models measuring the odds of having a mismatch between NDI and LexisNexis as the outcome. All models were adjusted for sex, marital status, race, Hispanic origin, Census Tract Poverty indicator, and age at diagnosis.
d. Numbers for these characteristics may not add to column totals due to small cell numbers.

e. An indicator for neighborhood poverty level based on the census tract of the diagnosis address(9).
f. Cells with <5 have been suppressed due to small numbers.

TABLE 2. Match Status between Residential State at Time of Death in LexisNexis Accurint and State of Death in National Death Index (NDI) among Cancer Recurrence and Information Surveillance (CRISP) Cohort Members who have died before and after manual review[a] (n=9,597)

Match Status	Before Manual Review, n (%)	After Manual Review, n (%)	Percent Change[b], %
NDI GA, LexisNexis GA	7,850 (81.8)	8,111 (84.5)	+3.2
NDI GA, LexisNexis Missing	738 (7.7)	738 (7.7)	0
NDI GA, LexisNexis Out of State	498 (5.2)	237 (2.5)	-110.1
NDI Out of State, LexisNexis GA	297 (3.1)	234 (2.4)	-26.6
NDI Out of State, LexisNexis Missing	36 (0.4)	36 (0.4)	0
NDI Out of State, LexisNexis Out of State, Match[c]	106 (1.1)	161 (1.7)	+34.2
NDI Out of State, LexisNexis Out of State, No Match[d]	27 (0.3)	35 (0.4)	+22.9
NDI Unknown, LexisNexis GA	36 (0.4)	36 (0.4)	0
NDI Unknown, LexisNexis Missing	<5	<5	0
NDI Unknown, LexisNexis Out of State	<5	<5	0
Sensitivity[e]	106/466 (23.7)	161/466 (34.6)	+31.5
Specificity[f]	7,850/9,086 (86.3)	8,111/9,086 (89.3)	+3.4

a. There were 795 cases (498 NDI said Georgia and LexisNexis said Outside of GA, and 297 NDI said Outside of GA and LexisNexis said GA) that were manually reviewed in the LexisNexis software to update the individual's residence to the most recent state of residence prior to their death.
b. Percent change was calculated as the number of cases before manual review subtracted from the number after manual review, divided by cases after manual review.
c. Refers to cases where both NDI and LexisNexis reported an out of state address, and the state in these cases were the same. These cases were included in the numerator when calculating sensitivity.
d. Refers to cases where both NDI and LexisNexis reported an out of state address, and the state in these cases were not the same. These cases were included in the denominator when calculating sensitivity.
e. Sensitivity was defined as the probability of being correctly classified as residing outside the state of Georgia in the LexisNexis database for those who died out of state. (NDI Out of State, LexisNexis Out of State, Match / Sum of all cases NDI reported as out of state)
f. Specificity was defined as the probability of being correctly classified as residing within the state of Georgia in the LexisNexis database for those who died in state. (NDI GA, LexisNexis GA/ Sum of all cases NDI reported as GA)

TABLE 3. Characteristics of the Cancer Recurrence and Information Surveillance Program (CRISP) who were not found in the LexisNexis Accurint database

Characteristic	**Missing in LexisNexis (n=3,604)** n (%)	**Found in LexisNexis (n=65,890)** n (%)	**Adjusted OR**[a] OR (95% CI)
Sex[b]			
Male	1,954 (54.2)	32,024 (48.6)	1.4 (1.3, 1.5)
Female	1,649 (45.8)	33,859 (51.4)	Reference
Marital status			
Single	1,240 (34.4)	8,474 (12.9)	4.0 (3.6, 4.3)
Married	1,151 (31.9)	37,543 (57.0)	Reference
Separated	73 (2.0)	704 (1.1)	2.6 (2.1, 3.4)
Divorced	339 (9.4)	6,457 (9.8)	1.8 (1.6, 2.0)
Widowed	383 (10.6)	6,191 (9.4)	2.8 (2.4, 3.2)
Unmarried/Domestic Partner	6 (0.2)	119 (0.2)	1.6 (0.7, 3.8)
Unknown	412 (11.4)	6,402 (9.7)	1.7 (1.5, 1.9)
Race			
White	1,761 (48.9)	43,260 (65.7)	Reference
Black	1,537 (42.7)	21,164 (32.1)	1.7 (1.7, 1.7)
American Indian/Alaskan Native	7 (0.2)	67 (0.1)	1.9 (1.9, 2.0)
Asian or Pacific Islander	258 (7.2)	1,245 (1.9)	6.9 (6.9, 7.0)
Other	5 (0.1)	8 (0.0)	-
Unknown	36 (1.0)	146 (0.2)	6.3 (6.3, 6.4)
Hispanic origin			
Non-Hispanic	3,153 (87.5)	63,979 (97.1)	Reference
Hispanic	451 (12.5)	1,911 (2.9)	5.5 (4.9, 6.1)
Census Poverty [b,c]			
0%-<5% Poverty	292 (8.1)	8,689 (13.2)	Reference
5%-<10% Poverty	557 (15.5)	13,728 (20.8)	1.2 (1.0, 1.4)
10%-<20% Poverty	1,139 (31.6)	22,406 (34.0)	1.5 (1.4, 1.8)
20%-100% Poverty	1,614 (44.8)	21,060 (32.0)	2.3 (2.1, 2.7)
Age at Diagnosis			
15-44	594 (16.5)	5,284 (8.0)	2.1 (1.9, 2.4)
45-54	537 (14.9)	10,940 (16.6)	1.0 (0.9, 1.1)
55-64	983 (27.3)	19,444 (29.5)	Reference
65-74	944 (26.2)	20,490 (31.1)	1.0 (0.9, 1.1)
75+	546 (15.2)	9,732 (14.8)	1.3 (1.2, 1.5)
Status			
Alive	2,824 (78.4)	57,073 (86.6)	Reference
Dead	780 (21.6)	8,817 (13.4)	2.1 (1.9, 2.3)

a. Multivariate logistic models measuring the odds of not having data in LexisNexis as the outcome. All models were adjusted for sex, marital status, race, Hispanic origin, census tract poverty indicator, age at diagnosis, and vital status.

b. Numbers for these characteristics may not add to column totals due to small cell numbers.
c.. An indicator for neighborhood poverty level based on the census tract of the diagnosis address(9)

TABLE 4. Characteristics of Cancer Recurrence of Information Surveillance Program (CRISP) Cohort Members by State of Residence according to LexisNexis Accurint[a]

Characteristic	Left GA (n=4,588) n (%)	Stayed in GA (n=61,302) n (%)	Adjusted OR[b] OR (95% CI)
Sex[c]			
Male	2,204 (48.0)	29,820 (48.6)	1.0 (0.9, 1,1)
Female	2,383 (51.9)	31,476 (51.4)	Reference
Marital status			
Single	708 (15.4)	7,766 (12.7)	1.4 (1.3, 1.5)
Married	2,380 (51.9)	35,163 (57.4)	Reference
Separated	65 (1.4)	639 (1.0)	1.6 (1.2, 2.1)
Divorced	531 (11.6)	5,926 (9.7)	1.4 (1.3, 1.6)
Widowed	446 (9.7)	5,745 (9.4)	1.3 (1.2, 1.5)
Unmarried/Domestic Partner	9 (0.2)	110 (0.2)	1.2 (0.6, 2.4)
Unknown	449 (9.8)	5,953 (9.7)	1.1 (1.0, 1.3)
Race[c]			
White	3,030 (66.0)	40,230 (65.6)	Reference
Black	1,421 (31.0)	19,743 (32.2)	1.0 (1.0, 1.1)
American Indian/Alaskan Native	13 (0.3)	54 (0.1)	2.8 (1.5, 5.1)
Asian or Pacific Islander	95 (2.1)	1,150 (1.9)	1.0 (0.8, 1.3)
Unknown	28 (0.6)	118 (0.2)	3.1 (3.1, 4.8)
Hispanic origin			
Non-Hispanic	4,357 (95.0)	59,622 (97.3)	Reference
Hispanic	231 (5.0)	1,680 (2.7)	1.8 (1.6, 2.1)
Census Poverty[c,d]			
0%-<5% Poverty	761 (16.6)	7,928 (12.9)	Reference
5%-<10% Poverty	1,127 (24.6)	12,601 (20.6)	0.9 (0.8, 1.0)
10%-<20% Poverty	1,501 (32.7)	20,906 (34.1)	0.7 (0.7, 0.8)
20%-100% Poverty	1,195 (26.1)	19,865 (32.4)	0.6 (0.6, 0.7)
Age at Diagnosis			
15-44	501 (10.9)	4,783 (7.8)	1.2 (1.0, 1.5)
45-54	770 (16.8)	10,170 (16.6)	0.9 (0.8, 1.1)
55-64	1,345 (29.3)	18,099 (29.5)	Reference
65-74	1,306 (28.5)	19,184 (31.3)	1.0 (0.9, 1.1)
75+	666 (14.5)	9,066 (14.8)	1.1 (0.9, 1.4)
Status			
Alive	3,954 (86.2)	53,119 (86.7)	Reference
Dead	634 (13.8)	8,183 (13.4)	1.1 (1.0, 1.3)

a. Excluding cases that were missing in LexisNexis Accurint.
b. Multivariate logistic models measuring the odds of not having data in LexisNexis as the outcome. All models were adjusted for sex, marital status, race, Hispanic origin, census tract poverty indicator, age at diagnosis, and vital status.
c. Numbers for these characteristics may not add to column totals due to small cell numbers.

d. An indicator for neighborhood poverty level based on the census tract of the diagnosis address(9)

FIGURE 1. Flow chart of Validation of LexisNexis using the Georgia Cancer Registry's Cancer Recurrence and Information Surveillance Program cohort