Residential history in cancer research: Utility of the annual billing ZIP code in the SEER-Medicare database and mobility among older women with breast cancer in the United States

S. Namin a,*, Y. Zhou b, E. McGinley b, K. Beyer b

a Institute for Health & Equity, Medical College of Wisconsin, Milwaukee, WI, USA
b Center for Advancing Population Science, Medical College of Wisconsin, Milwaukee, WI, USA

A R T I C L E   I N F O

Keywords:
- Cancer
- Residential history
- Residential mobility
- SEER-Medicare linked database

A B S T R A C T

There is a rise in attention to residential history in cancer epidemiology aimed at more effective estimation of social and physical environmental exposures and the influence of place of residence on cancer outcomes. However, in the United States, as in many other countries, residential history data are not readily available. In this paper we explore the feasibility of using the annual Medicare billing ZIP code history available in the SEER-Medicare database to study residential mobility among older cancer survivors in the U.S. In a cohort of women diagnosed with breast cancer between 2007 and 2015, we examine the completeness of the data along with the overall characteristics of residential moves based on race and stage at diagnosis. Findings indicate that residential mobility among older women with breast cancer in the U.S. is limited, but differences by race/ethnicity, stage at diagnosis and before/after diagnosis are statistically significant. And breast cancer survivors from minority groups move more frequently than their non-Hispanic White counterparts. The results also show that move rate slightly, but statistically significantly, increases after diagnosis. We conclude that SEER-Medicare can be utilized to study residential mobility among older cancer survivors. We recommend the creation of subcohorts based on specific research questions to account for variability in residential mobility due to very short survival times or a diagnosis shortly after Medicare enrollment. Studying residential history provides the opportunity for assigning socioecological and exposure metrics for future survival studies.

Introduction

There is a rise in attention to residential history in cancer epidemiology to improve exposure assessment and enable detailed studies of housing and mobility on cancer outcomes (Stinchcomb & Roesser, 2016). Characterizing residence over time is particularly important when considering significant latency periods between exposure and diagnosis for many cancers. For example, a 2014 study showed that many cancers have more than 10 years of latency (Nadler, Zurbenko, & Buchanich, 2014). However, in many countries, including the U.S., residential history data are not readily available. In fact, only a few countries have nationwide databases that provide life-time residential histories (Nikkila et al., 2018). Therefore, residential location at the time of diagnosis is often used to operationalize studies of neighborhood effects and physical environment exposures (Timander & McLafferty, 1998). The reliance on these static measures is further reinforced by the time and cost of obtaining residential history information.

Some researchers choose to establish residential history datasets based on self-reported residential histories through in-person interviews and questionnaires (Gallagher, Webster, Aschengrau, & Vieira, 2010; Little et al., 2018; Nordsborg et al., 2015; Urayama et al., 2009; Wu et al., 2014). Recent studies have looked at the reliability of self-reported residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014). Moreover, in the U.S, some studies have evaluated the use of commercially available credit reporting companies such as LexisNexis to construct residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014). Moreover, in the U.S, some studies have evaluated the use of commercially available credit reporting companies such as LexisNexis to construct residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014). Moreover, in the U.S, some studies have evaluated the use of commercially available credit reporting companies such as LexisNexis to construct residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014). Moreover, in the U.S, some studies have evaluated the use of commercially available credit reporting companies such as LexisNexis to construct residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014). Moreover, in the U.S, some studies have evaluated the use of commercially available credit reporting companies such as LexisNexis to construct residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014). Moreover, in the U.S, some studies have evaluated the use of commercially available credit reporting companies such as LexisNexis to construct residential histories for epidemiological studies and have concluded in favor of these approaches when no other source is available (Drozdovitch et al., 2016; Sonderman, Tarone, & McLaughlin, 2014).
though this detailed information is costly (Jacquez et al., 2011).

The SEER-Medicare database is a linkage of data from the Surveillance, Epidemiology, and End Results (SEER) cancer registries in the U.S. with billing claims from Medicare. It is a valuable tool often leveraged for conducting cancer research in elderly populations. There are currently over 2.2 million cases in the dataset from 1991 to 2015 (National Cancer Institute, 2019). Interest in utilizing this dataset to study cancer screening, treatment and use of health care resources, survival, and cancer care costs is growing rapidly (Beebe-Dimmer et al., 2019; Brar et al., 2019; Rosenblatt, Osterbur, & Douglas, 2016; Schwartz et al., 2018). Moreover, recent studies have employed this dataset in order to study disparities in cancer outcomes across different racial groups as well as geographic variations in cancer care (Beebe-Dimmer et al., 2019; Lam, Cronin, Ballard, & Mariotto, 2018; Ratnapradipa et al., 2017; Suzuki, Callen, Mehra, Bentzen, & Goloubeva, 2019; Williams et al., 2019). Additional studies have looked at opportunities and limitations of using SEER-Medicare to assess environmental exposure and to conduct case control studies (Engels et al., 2011; VoPham et al., 2015).

In addition to cancer site, stage, initial treatment, patient demographics, vital status, and claims for all covered fee-for-service health care from the time of enrollment in Medicare to death, the SEER-Medicare database also includes annual billing ZIP Codes starting at the time of enrollment in Medicare. While some studies utilizing SEER-Medicare use Census data at the census tract and ZIP code level to estimate neighborhood socioeconomic status (SES) characteristics, the ZIP Code history provided in this dataset may provide an additional opportunity for examining the nature and characteristics of residential movement and related exposures. Although there are studies that have used ZIP Code at the time of diagnosis or death, obtained from SEER-Medicare database, mainly to calculate SES status or access to healthcare (Adams et al., 2017; Charlton et al., 2016; Jarosek, Shippee, & Virnig, 2016), to our knowledge no studies have examined the feasibility of using the Medicare billing ZIP Codes contained in the SEER-Medicare database to study residential histories among older cancer survivors, despite the minimal cost and wide coverage of the database.

Given growing attention to residential history in cancer research and the difficulty of access to this information in the U.S., SEER-Medicare data provide a potential opportunity for studying residential movements among older cancer survivors across the country at minimal cost. The primary objective of this study is to determine the feasibility of using reported Medicare billing ZIP Codes in the SEER-Medicare dataset to study residential movement among older women diagnosed with breast cancer in the U.S. To assess the feasibility of leveraging this data to study residential movements, we address the following questions in a contemporary cohort of older women diagnosed with breast cancer.

1) Does SEER-Medicare contain complete ZIP code histories before and after diagnosis?
2) How often do women move? Do they move more or less often after diagnosis?
3) How many unique ZIP Codes (neighborhood environments) do women experience?
4) Does the move rate or number of unique ZIP codes vary by race/ethnicity or stage of disease at diagnosis?

Methods

Data. We used the SEER-Medicare dataset to examine a cohort of women diagnosed with breast cancer between 2007 and 2015 with follow up until 2017 (N = 106,565 across 17 SEER registries). The cohort included women who were diagnosed with their first breast cancer in 2007–2015, were age 66–84 years and alive at diagnosis, had a ZIP code and resided in a Metropolitan Statistical Area (MSA) at diagnosis. Since most studies using SEER-Medicare data are interested in calculating comorbidity, we adopted a cohort age 66 and older at diagnosis to ensure Medicare enrollment one year prior to diagnosis for the calculation of comorbidity. Non-DCIS Stage 0 cases were excluded. Special permission from SEER-Medicare was obtained in order to receive and analyze billing ZIP code for each year enrolled in Medicare. Annual ZIP code data reflect the ZIP code for the billing address at the end of each calendar year enrolled. Beneficiary race/ethnicity and stage at diagnosis were obtained from the SEER registry. Among the full cohort, there were <11 cases of women for whom ZIP code was missing for several years. We imputed the missing ZIP Codes for these cases based on the rationale that if the recorded ZIP Codes before and after the missing record were the same, the patient had not moved. We also constructed a subset of the cohort consisting of women with at least 3 years of known residential ZIP code records both before and after diagnosis (n = 74,722). This allowed for comparison of pre and post diagnosis periods while removing cases with short survival and/or short pre-diagnosis Medicare enrollment periods.

Measures. We created two measures related to residential moves: (1) number of moves per year (move rate) and (2) number of unique ZIP Codes. The move rate reflects the residential mobility of the cohort. The number of unique ZIP Codes is indicative of the number of different environments to which a patient has been exposed. For example, in some cases a patient moved back to a ZIP Code after living in another ZIP Code for a period of time. For these cases, we considered the repeated ZIP Code in calculating the move rate, but each ZIP Code is counted only once in calculating the number of unique ZIP Codes.

Analysis. Descriptive statistics were used to summarize the data, including race and ethnicity, age at diagnosis, stage at diagnosis, and duration of Medicare enrollment. We examined the overall move rate and number of unique residential ZIP codes, including before and after diagnosis separately. Pre and post diagnosis periods did not include the ZIP code of residence in the diagnosis year. We performed the Wilcoxon signed rank test to compare the move rate and the number of unique ZIP Codes pre and post diagnosis. We used the Kruskal-Wallis test and Wilcoxon test to analyze the statistical significance of race and stage of diagnosis for both move rate and the number of unique ZIP codes since enrollment.

Results

All results are presented in Table 1, which summarizes both the full cohort (N = 106,565) and the sub-cohort of women with at least 3 years of both pre and post diagnosis time (N = 74,722).

Summary of the cohorts

The majority of both cohorts were White (84.4% in full cohort; 84.9% in sub-cohort). The majority were diagnosed between the ages of 75–84 (38.6% in full cohort; 41.9% in sub-cohort). Also, in both cohorts the majority were diagnosed at stage I (43.3% in full cohort; 47.0% in the sub-cohort). The average number of years of Medicare enrollment prior to diagnosis was longer for both cohorts than the survival time post diagnosis, with 10.76 years and 11.13 years pre diagnosis and 4.90 years and 5.91 years post diagnosis, for the full and sub-cohorts respectively. The complete breakdowns of both cohorts are presented in Table 1, Section 1.

Question 1: Does SEER-Medicare contain complete ZIP code histories before and after diagnosis?

Complete ZIP Code histories were available for 99.96% of all the patients from the time of first breast cancer diagnosis in 2007–2015 until the time of death, or for surviving patients until 2017. It should be noted that even before imputation (explained in the methods), the percentage with complete residential histories after diagnosis was over 99%. Data are presented in Table 1, Section 1.

Question 2: How often do women move? Do they move more or less often after diagnosis?

Moves and move rates are shown in Table 1, Section 2. Overall, both cohorts have low residential mobility. Examining the full cohort, the
average move rate is only 0.03 moves per year; 60.67% of women did not move, while less than 7% moved 3 or more times. Similar numbers are observed for the sub-cohort. In the sub-cohort, the average move rate is 0.06, 58.44% did not move and 7.28% moved 3 or more times. This average move rate during the period of observation in SEER Medicare would translate to, on average, a move every 30 years.

To compare the move rate pre and post diagnosis, we examine the sub-cohort to ensure data for at least 3 years pre and 3 years post diagnosis. The results show that move rate slightly, but statistically significantly, increases after diagnosis (from 0.041 to 0.044 moves per year).

Question 3: How many unique ZIP Codes (neighborhood

Table 1
Characteristics of the full cohort (N = 106,565) and the sub-cohort (N = 74,722).

| Section 1: Summary Characteristics | Full Cohort (n = 106,565) | Sub-Cohort (n = 74,722) |
|------------------------------------|---------------------------|-------------------------|
| **Age at Diagnosis**               |                           |                         |
| 66-69                              | 24748 (23.23%)            | 12990 (17.39%)          |
| 70-74                              | 28478 (26.72%)            | 22844 (30.57%)          |
| 75-84                              | 41223 (38.68%)            | 31316 (41.91%)          |
| 85 or older                        | 12116 (11.37%)            | 7572 (10.13%)           |
| **Race/Ethnicity**                 |                           |                         |
| White                              | 89937 (84.40%)            | 63508 (84.99%)          |
| African-American/Black             | 9506 (8.64%)              | 6021 (8.06%)            |
| Asian                              | 2985 (2.80%)              | 2150 (2.88%)            |
| Hispanic                           | 1346 (1.26%)              | 938 (1.25%)             |
| Native American                    | 233 (0.2%)                | 154 (0.21%)             |
| Others                             | 2858 (2.68%)              | 1951 (2.61%)            |
| **Stage at Diagnosis**             |                           |                         |
| 0                                  | 16127 (15.13%)            | 12415 (16.61%)          |
| 1                                  | 46235 (43.39%)            | 35178 (47.08%)          |
| 2                                  | 26549 (24.91%)            | 18607 (24.90%)          |
| 3                                  | 7732 (7.26%)              | 4674 (6.26%)            |
| 4                                  | 5404 (5.07%)              | 1504 (2.01%)            |
| **Number of years of Medicare enrollment** |                   |                         |
| Overall                            | 16.66 (9.07%)             | 18.04 (5.91%)           |
| **Pre-diagnosis**                  | 10.76 (5.97%)             | 11.13 (5.35%)           |
| **Post-diagnosis**                 | 4.90 (2.73%)              | 5.91 (2.22%)            |
| **Records with Complete ZIP Codes Since Enrollment** |       |                         |
| Overall                            | 106529 (99.96%)           | 74722 (100)             |

| Section 2: Moves and Move Rate     |                           |                         |
| **Number of moves since enrollment** |                           |                         |
| 0                                  | 64655 (60.67%)            | 43665 (58.44%)          |
| 1-2                                | 34703 (32.56%)            | 25612 (34.27%)          |
| 3+                                 | 7207 (6.76%)              | 5445 (7.28%)            |
| **Number of moves pre-diagnosis**  |                           |                         |
| 0                                  | 75326 (70.68%)            | 52382 (70.10%)          |
| 1-2                                | 27020 (25.35%)            | 19551 (26.16%)          |
| 3+                                 | 4212 (3.95%)              | 2799 (3.73%)            |
| **Number of moves post-diagnosis** |                           |                         |
| 0                                  | 88189 (82.75%)            | 59306 (79.36%)          |
| 1-2                                | 17700 (16.60%)            | 14805 (19.81%)          |
| 3+                                 | 676 (0.63%)               | 611 (0.81%)             |
| **Number of moves per year (move rate)** |                 |                         |
| Overall                            | 0.03 (0.06%)              | 0.03 (0.06%)            |
| Pre-diagnosis                      | 0.07 (0.13%)              | 0.04*** (0.08)          |
| Post-diagnosis                     | 0.04 (0.12%)              | 0.04** (0.10)           |

| Section 3: Unique ZIP codes        |                           |                         |
| **Number of unique ZIP codes since enrollment** |         |                         |
| 1                                  | 62164 (58.33%)            | 42050 (56.28%)          |
| 2                                  | 28094 (26.36%)            | 20433 (27.35%)          |
| 3+                                 | 16308 (15.30%)            | 12239 (16.37%)          |
| **Number of unique ZIP codes pre-diagnosis** |             |                         |
| 1                                  | 77702 (72.91%)            | 53803 (72.00%)          |
| 2                                  | 20309 (19.05)             | 15001 (20.07)           |
| 3+                                 | 8548 (8.02)               | 5918 (7.92)             |
| **Number of unique ZIP codes post-diagnosis** |          |                         |
| 0**                                | 3747 (3.51)               | 0 (0)                   |
| 1                                  | 87531 (82.13)             | 61284 (82.01)           |
| 2                                  | 13189 (12.37)             | 11518 (15.41)           |
| 3+                                 | 2099 (1.96)               | 1920 (2.56)             |

Signif. Codes: ***$<0.001$, **$<0.01$, *$<0.05$.
1 Wilcoxon rank-sum test was performed to compare the move rate before and after diagnosis for the sub-cohort.
2 In the Full cohort 3747 of the women had survival less than a year and therefore no ZIP Code is recorded after the year of diagnosis.

To compare the move rate pre and post diagnosis, we examine the sub-cohort to ensure data for at least 3 years pre and 3 years post diagnosis. The results show that move rate slightly, but statistically significantly, increases after diagnosis (from 0.041 to 0.044 moves per year).

Question 3: How many unique ZIP Codes (neighborhood

Panel (a): move rate by race/ethnicity

Panel (b): move rate by stage of disease at diagnosis

Fig. 1. Comparison of move rate by (a) race/ethnicity and (b) stage of disease at diagnosis at 95% confidence interval.
importance of subjective rather than objective measures of neighbor
tients in the U.S. This near-complete availability of ZIP code level
over, analyzing residential history among cancer survivors before and
dential stability that is associated with health outcomes.
history analysis in this dataset.
allow for analysis for movements inside the same ZIP Code and relying
matched ZIP Codes. Moreover previous studies show that census tract
used geographical unit to report heath and socioecological surveillance
Limitations
Codes based on stage were statistically significant at 95% confidence
known stage of diagnosis experience more ZIP Codes, on average 1.86,
and block group result in different gradients in SES compared with ZIP
Codes for women in the full cohort was only 1.66.
Breast cancer survivors diagnosed at stages II, III and unknown
move more frequently (Fig. 1(b)). Overall, compared to stage 0,
the differences in move rate were statistically significant for stage II, III,
and unknown stage at 95% confidence interval. Also, women with un-
known stage of diagnosis experience more ZIP Codes, on average 1.86,
comparing to women with known stages. And the number of unique ZIP
Codes based on stage were statistically significant at 95% confidence
interval.

Limitations

Although ZIP Code has significant limitations it remains a widely
geographical unit to report heath and sociocological surveillance
data due to cost effectiveness. One of the limitations of adopting ZIP
Code for temporal analysis is that the boundaries are not precise or
stable and do not always match ZIP Code Tabulation Area (ZCTA) that is
used by the Census Bureau (Beyer, Schultz, & Rushton, 2008; Krieger
et al., 2002). This will require an additional step in marking the un-
matched ZIP Codes. Moreover previous studies show that census tract
and block group result in different gradients in SES compared with ZIP
Code (Krieger et al., 2002). Recent research also emphasizes the
importance of subjective rather than objective measures of neighbor-
hood conditions that can affect residential mobility (Jones & Dantzler,
2020) and health outcomes.

In addition to these limitations, SEER-Medicare dataset does not
allow for analysis for movements inside the same ZIP Code and relying
on unique ZIP Codes is the only way to operationalize a residential
history analysis in this dataset.

Despite these limitations, given that many health estimates and
environmental measures in the U.S. are available at ZIP Code level,
utilizing annual Zip codes reported by the SEER-Medicare can advance
research on the effects of neighborhood context on cancer survivorship.
Understanding residential mobility of cancer survivors is important for
analyzing sociocological environment and it is also a measure of resi-
dential stability that is associated with health outcomes.

Discussion

The construction of residential histories is an important part of
epidemiological studies, especially for diseases with long latencies such
as cancer. Linking residential history data with historical environmental
data to calculate lifetime exposure can help identify risk agents. More-
over, analyzing residential history among cancer survivors before and
after their diagnosis can shed light on people’s experiences of living with
cancer. Given the cost and time constraints of other methods of
obtaining residential history and the 99% complete annual ZIP code data
in the database, SEER-Medicare can be considered an underutilized
resource of information on residential history amongst older cancer pa-
ents in the U.S. This near-complete availability of ZIP code level
residential location information on an annual basis could be leveraged to
enhance social and physical environmental exposure assessment and
explore the impact of residential mobility in breast cancer outcomes
among older women in the US.

Residential mobility overall was low among this cohort, and no
meaningful differences, despite statistical significance were observed
pre and post diagnosis. On average women in the sub-cohort moved
every 30 years. However, some differences were observed with regard to
stage and race/ethnicity, which could prove useful in examining dis-
parities in breast cancer outcomes. Differences in move rate based on
race and ethnicity echoes the overall mobility characteristic among the
general population reported by the Census mobility data (US Census
Bureau, 2019) and epidemiological studies of older women (Medgyesi
et al., 2020). However, since residential history data from the
SEER-Medicare dataset starts after Medicare enrollment and survival
can vary significantly, we suggest that studies create clearly defined
sub-cohorts based on the specific research question at hand, in order to
account for variability in residential mobility due to very short survival
or a diagnosis in the first or second year of Medicare enrollment.
Additionally, such studies will of course inherit the general limitation of
the application of ZIP code as the geographic unit for analysis instead of
smaller or more precise geographies, including individual addresses,
census block groups, and census tracts.

Future research should explore this dataset to study residential
mobility as it relates to other aspects of cancer care, such as treatment
and survival, as well as additional cancer sites beyond breast cancers.

Declaration of competing interest

Authors declare no conflict of interests.

Acknowledgements

This study used the linked SEER-Medicare database. The interpre-
tation and reporting of these data are the sole responsibility of the au-
thors. The authors acknowledge the efforts of the National Cancer
Institute; the Office of Research, Development and Information, CMS;
Information Management Services (IMS), Inc.; and the Surveillance,
Epidemiology, and End Results (SEER) Program tumor registries in the
creation of the SEER-Medicare database.

The collection of cancer incidence data used in this study was sup-
ported by the California Department of Public Health as part of the
statewide cancer reporting program mandated by California Health and
Safety Code Section 103885; the National Cancer Institute’s Surveil-
ance, Epidemiology and End Results Program under contract
HHSN26120100140C awarded to the Cancer Prevention Institute of
California, contract HHSN26120100035C awarded to the University of
Southern California, and contract HHSN26120100034C awarded to
the Public Health Institute; and the Centers for Disease Control and
Prevention’s National Program of Cancer Registries, under agreement
#US8PD003862-01 awarded to the California Department of Public
Health. The ideas and opinions expressed herein are those of the author
(s) and endorsement by the State of California Department of Public
Health, the National Cancer Institute, and the Centers for Disease Con-
trol and Prevention or their Contractors and Subcontractors is not
intended nor should be inferred.

This work was supported by grants from the National Institutes of
Health (NCI R01CA214805, T32HP10030). The content is solely the
responsibility of the authors and does not necessarily represent the
official views of the National Institutes of Health or the Health Resources
and Services Administration (HRSA).

We thank Courtney Jankowski for reviewing the paper and providing
feedback.
