Residential Mobility and Lung Cancer Risk: Data-Driven Exploration Using Internet Sources

Hong-Jun Yoon[1], Georgia Tourassi[1], and Songhua Xu[2]

[1] Health Data Sciences Institute, Oak Ridge National Laboratory
Oak Ridge, TN 37831, United States
{yoonh,tourassig}@ornl.gov
[2] Department of Information Systems, College of Computing Sciences,
New Jersey Institute of Technology, University Heights, Newark, NJ, 07102, United States
songhua.xu@njit.edu

Abstract. Frequent relocation has been linked to health decline, particularly with respect to emotional and psychological wellbeing. In this paper we investigate whether there is an association between frequent relocation and lung cancer risk. For the initial investigation we used web crawling and tailored text mining to collect cancer and control subjects from online data sources. One data source includes online obituaries. The second data source includes augmented LinkedIn profiles. For each data source, the subjects' spatiotemporal history is reconstructed from the available information provided in the obituaries and from the education and work experience provided in the LinkedIn profiles. The study shows that lung cancer subjects have higher mobility frequency than the control group. This trend is consistent for both data sources.

Keywords. Residential Mobility, Lung Cancer, Social Media, Health Data Informatics

1 Introduction

There is rich literature in life-course epidemiology investigating the relationship between residential mobility and a person's health. The studies indicate an adverse effect and a fairly complex relationship, which includes both social and environmental factors. A systematic review of twenty-two studies [1] from the medical and social sciences literature reported that frequent residential change during childhood is a clinical risk marker of behavioral and emotional health. Tønnessen et al [2] showed that frequent relocation during the early adolescent years is more detrimental than reloca-

tion in early childhood. Lin [3] studied the relationship between residential mobility history and self-rated health at midlife observing a similar association.

The association between residential mobility and cancer has been studied only in the context of radon exposure and lung cancer [e.g., 4-5]. The purpose of this paper is to explore the potential of a novel cyber-informatics approach to study a similar relationship, specifically if lifetime residential mobility is associated with lung cancer risk. Our study focuses on lung cancer because it is the leading cause of cancer death in the United States [6] both for males and females.

2 Methods

2.1 Data Sources

The typical retrospective case-control observational study involves first the collection of subjects with and without a specific disease (i.e., lung cancer) and then identification of each subject's exposure to the specific condition under investigation (i.e., lifetime residential mobility). We identified two different online data sources that contain the basic information needed for our study in a form that is relatively easily interpretable by computer. These sources include online obituaries and augmented LinkedIn social network profiles.

2.1.1 Online Obituaries

Online obituaries are widely available in newspaper sites, funeral homes' web pages, and web-based obituary archives. They have a largely similar format consisting of four sections; death announcement, biographical information, survivor information, and information about the funeral arrangements. Intrinsically, obituaries include basic information of the deceased that is essential for our study; name, age, birth date, cause of death, residence, and sometimes major life events (e.g., schools attended, military service, employments). Using an advanced web crawler developed in our laboratory [7], we searched the Internet for obituaries of people who died of lung cancer as well as non-cancer related obituaries. We applied the CoreNLP software package [8] to understand the obituaries' text content. Obituaries from which we could not identify or infer the essential information (i.e., age and gender) were excluded from this study.

- *Name and Age:* Typically a subject's name and age at death is explicitly stated in the first sentence (e.g.: "John Doe, 63, of Oak Ridge, TN passed away May 31, 2013."). In other cases age is inferred from the content by detecting dates at birth and at death (e.g.: "John Doe passed away Sunday, December 31, 2013… He was born January 1, 1950…").
- *Gender:* Gender is inferred by calculating the prevalence of male and female pronounces present in the obituary (e.g.: "She passed away at her residence…").
- *Cause of Death:* Lung cancer history was inferred from explicit statements (e.g.: "He passed away after a courageous battle with lung cancer…"). We applied heuristic rules to filter out those obituaries that may contribute to false counts. For example, obituaries including sentences stating the family prefers monetary contribu-

tions to cancer research foundation rather than flowers (e.g.: "In lieu of flowers, please consider donations to lung cancer research.") were not considered cancer cases.

- *Locations of Residence:* Locations of residence of the deceased are stated in content, usually city and state. We collect locations of residence from the birthplace, residences, and the address where the funeral took place.

2.1.2 Augmented LinkedIn Dataset

LinkedIn is a business-oriented social network service which includes professional profiles (i.e., job history, work experience and education). Although LinkedIn profiles are a rich source of subjects with detailed spatiotemporal information during adulthood, few LinkedIn profiles contain the subjects' medical information, for example whether they have battled cancer. To leverage the advantages of the LinkedIn profiles while mitigating the limitations, we developed an additional cyber-informatics step.

Stories of lung cancer patients and survivors are abundant on the Internet, such as in open cancer survivor networks, lung cancer survivors' blogs, as well as national and local newspapers presenting lung cancer survivors' stories. First, we crawled such life stories of lung cancer patients and survivors using the advanced web crawler mentioned earlier. From the life stories of those candidate cancer subjects, we identified the subjects' names and other information that helped us search and match them with profiles available in LinkedIn. Since LinkedIn profiles have no direct indication of the subjects' gender and age, we developed tailored algorithms to infer this information. Gender was inferred from the first name of the LinkedIn profile utilizing the genderize.io API [9]. Education history (i.e., high school graduation year and college years) enabled us to estimate the subject's age. Profiles for which gender and age could not be inferred were excluded from further analysis. Furthermore, subjects with LinkedIn profiles with less 10 years location history were excluded from further analysis.

2.2 Number of Relocations in Lifetime

Collected residence locations from an obituary were aligned chronologically and then converted into geographical codes. Residence locations in LinkedIn profiles were collected from the "Experiences" and "Education" sections from which chronological residence locations were recomposed. We applied a simple rule to determine whether a particular geographical move was significant. If the distance of two consecutive locations was less than 50 miles, it was not considered a major move and was not included in the calculation of a person's mobility history.

3 Results

3.1. Online Obituary Dataset

Following the procedure described in Section 2.1 we formed a case group with lung cancer subjects and a control group of cancer-free subjects. To replicate a matched case-control study design, age and gender adjustment was achieved by selecting the same number of case and control subjects for each age and gender group.

The total number of lung cancer obituaries was 27,391. We found more male lung cancer subjects (16,129) than females (11,262). Table 1 shows the number of relocations per gender and age group. Statistical comparisons were made using the Student's t-test. Overall the average number of relocations for lung cancer subjects was 4.09±3.16, which is significantly higher than that for the lung-cancer free group 3.01±2.65. Significantly higher mobility was observed for the case group than the control group for all age groups and both genders consistently.

Table 1. Obituary Dataset: Number of obituaries of lung cancer diseased and cancer-free subjects by age and gender, average number of relocations (ARL), and their statistical comparison.

GENDER	AGE GROUP	NO. OF SUBJECTS	ARL CASES	ARL CONTROLS	P-VALUE
Female	**All**	**11,262**	**4.45**	**3.13**	**< 1e-5**
	20~29	114	5.31	2.91	< 1e-5
	30~39	169	4.75	3.37	0.002
	40~49	580	3.74	2.83	< 1e-5
	50~59	1,686	3.77	2.73	< 1e-5
	60~69	2,886	4.08	3.08	< 1e-5
	70~79	3,471	4.46	3.20	< 1e-5
	80~89	2,048	5.30	3.41	< 1e-5
	90~	308	6.56	3.59	< 1e-5
Male	**All**	**16,129**	**3.84**	**2.92**	**< 1e-5**
	20~29	143	4.27	1.85	< 1e-5
	30~39	199	3.85	2.50	< 1e-5
	40~49	738	3.14	2.15	< 1e-5
	50~59	2,504	3.00	2.39	< 1e-5
	60~69	4,451	3.47	2.83	< 1e-5
	70~79	5,050	4.01	3.07	< 1e-5
	80~89	2,703	4.83	3.50	< 1e-5
	90~	341	6.00	3.60	< 1e-5
All	**All**	**27,391**	**4.09**	**3.01**	**< 1e-5**

Table 2. Obituary Dataset: Odds ratios (OR) and 95% confidence intervals (CI) between low mobility and high mobility groups, stratified by age and gender.

GENDER	FEMALE		MALE	
AGE GROUP	NO. OF SUBJECTS	OR (95% CI)	NO. OF SUBJECTS	OR (95% CI)
All	**11,262**	**1.75 (1.66~1.85)**	**16,129**	**1.47 (1.41~1.54)**
20~29	114	3.03 (1.75~5.24)	143	3.65 (2.23~5.97)
30~39	169	2.17 (1.40~3.35)	199	2.17 (1.45~3.24)
40~49	580	1.47 (1.17~1.86)	738	1.72 (1.40~2.13)
50~59	1,686	1.74 (1.52~2.00)	2,504	1.40 (1.25~1.57)
60~69	2,886	1.57 (1.42~1.75)	4,451	1.38 (1.27~1.50)
70~79	3,471	1.74 (1.58~1.91)	5,050	1.47 (1.36~1.59)
80~89	2,048	2.00 (1.76~2.27)	2,703	1.55 (1.39~1.73)
90~	308	3.11 (2.20~4.38)	341	1.95 (1.43~2.66)
All	**27,391**		**1.58 (1.53~1.64)**	

We also measured the odds ratio (OR) between low mobility (0~2 times of relocation) and high mobility (3+ times of relocation) subjects (Table 2). We observed higher lung cancer incidence in high mobility subjects for all age groups and genders. Interestingly the OR is higher in younger groups for both genders.

3.2. Augmented LinkedIn Dataset

We repeated the same statistical analysis with subjects' profiles from the augmented LinkedIn dataset (Table 3). To replicate a matched case-control study design, we randomly selected an equal number of LinkedIn profiles with similar age and gender distribution from the remaining LinkedIn population.

Table 3. Augmented LinkedIn Dataset: Number of obituaries of lung cancer and cancer-free subjects by age and gender, average number of relocations (ARL), and statistical comparison.

GENDER	AGE GROUP	NO. OF SUBJECTS	ARL CASES	ARL CONTROLS	P-VALUE
Female	**All**	**143**	**3.04**	**2.85**	**0.342**
	20~29	33	2.94	2.59	0.321
	30~39	45	2.87	2.99	0.732
	40~49	33	3.06	2.96	0.807
	50~59	26	3.23	2.90	0.522
	60~69	6	4.00	2.57	0.134
Male	**All**	**207**	**3.16**	**2.83**	**0.048**
	20~29	33	3.15	3.04	0.817
	30~39	66	3.29	2.61	0.019
	40~49	46	3.17	2.97	0.555
	50~59	33	3.09	3.08	0.983
	60~69	25	2.96	2.62	0.427
All	**All**	**350**	**3.11**	**2.84**	**0.032**

Table 4. Augmented LinkedIn Dataset: Odds ratios (OR) and 95% confidence intervals (CI) between low mobility and high mobility groups, stratified by age and gender.

GENDER	FEMALE		MALE	
AGE GROUP	NO. OF SUBJECTS	OR (95% CI)	NO. OF SUBJECTS	OR (95% CI)
All	**143**	**1.06 (0.66~1.69)**	**207**	**1.10 (0.75~1.63)**
20~29	33	2.92 (1.06~8.03)	33	0.95 (0.36~2.51)
30~39	45	0.48 (0.20~1.12)	66	1.21 (0.60~2.41)
40~49	33	0.89 (0.34~2.33)	46	1.43 (0.62~3.26)
50~59	26	1.36 (0.46~4.05)	33	0.88 (0.33~2.34)
60~69	6	2.00 (0.19~20.61)	25	0.85 (0.28~2.58)
All	**350**		**1.08 (0.80~1.46)**	

It was observed that the average number of relocations in the case group was significantly higher than the average number of relocations in the control group (3.04±1.75 vs. 2.85±1.55). For each age and gender group, we observed mostly higher mobility in the cases than the controls. However, most differences were not statistical-

ly significant, possibly due to the smaller sample size of each subgroup. We also calculated the odds ratio of lung cancer risk between low mobility and high mobility groups observing similar trends but no significant differences within age and gender subgroups (Table 4).

4 Discussion

We presented a cyber-informatics approach to study the relationship between lifetime residential mobility frequency and lung cancer risk. The study utilized two distinct non-traditional data cybersources, namely obituaries and LinkedIn and replicated a matched case-control retrospective study design. Each data source shared its own strengths and limitations in terms of the sampling biases introduced. Regardless of the distinct differences between the two data sources, we observed consistently that frequent (and substantial) geographical relocation is linked to higher lung cancer risk. Furthermore, this study illustrated how non-traditional big data sources can be leveraged to execute cost-effective in silico epidemiological studies for knowledge discovery and hypotheses generation. However, issues of sampling bias and techniques to mitigate these challenges are still work in progress.

Acknowledgements

The study was funded by NIH/NCI (Grant #: 1R01CA170508-03).

References

1. Jelleyman, T., Spencer, N.: Residential Mobility in Childhood and Health Outcomes: A Systematic Review. J. Epidemiol Community Health. 62, 584-592 (2007)
2. Tønnessen, M., Telle, K., Syse, A.: Childhood Residential Mobility and Adult Outcomes. Statistics Norway Research Department. Discussion Papers. No. 750 (2013)
3. Lin, K. C., Huang, H. C., Bai, Y. M., Kuo, P. C.: Lifetime residential mobility history and self-rated health at midlife. Journal of Epidemiology, 22(2), 113-122 (2012)
4. Warner, K. E., Mendez, D., Courant, P. N.: Toward a more realistic appraisal of the lung cancer risk from radon: the effects of residential mobility. American Journal of Public Health, 86(9), 1222-1227 (1996)
5. Krewski, Daniel, et al.: Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. Epidemiology, 16(2), 137-145 (2005)
6. American Cancer Society: Cancer facts & figures. (2014)
7. Xu, S., Yoon, H. J., Tourassi, G. D.: A user-oriented web crawler for selectively acquiring online content in e-health research. Bioinformatics, 30(1), 104-114 (2014)
8. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60 (2014)
9. Determine the Gender of a First Name, http://genderize.io/#overview