HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2016 October 04.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2016 February ; 2016: 557–560. doi:10.1109/BHI.

2016.7455958.

Investigating the Association Between Sociodemographic Factors and Lung Cancer Risk Using Cyber Informatics

Hong-Jun Yoon and

Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (phone: 865-241-2626; fax: 865-574-6275; yoonh@ornl.gov).

Georgia Tourassi [Member, IEEE]

Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (phone: 865-576-4829; fax: 865-574-6275; tourassig@ornl.gov).

Abstract

Openly available online sources can be very valuable for executing in silico case-control epidemiological studies. Adjustment of confounding factors to isolate the association between an observing factor and disease is essential for such studies. However, such information is not always readily available online. This paper suggests natural language processing methods for extracting socio-demographic information from content openly available online. Feasibility of the suggested method is demonstrated by performing a case-control study focusing on the association between age, gender, and income level and lung cancer risk. The study shows stronger association between older age and lower socioeconomic status and higher lung cancer risk, which is consistent with the findings reported in traditional cancer epidemiology studies.

I. Introduction

Case-control studies are widely used to discover risk factors in cancer epidemiology. Traditional case-control studies require tremendous effort to recruit participants and collect detailed data from them. In our previous work we have shown how to leverage cyber-informatics for in silico epidemiological discoveries using information that people share openly online [1]. Our approach is an example of the recently established field known as digital epidemiology. Although fully automated and cost-effective, digital epidemiology has limitations due to the variable amount and granularity of information online users share voluntarily. Information related to age, gender, and socioeconomic status is not always readily available, yet all three are factors are very important due to biases they may introduce skewing the study findings. To mitigate the risk of bias, traditional epidemiological studies use detailed questionnaires to collect every potential confounding factor. With online sources though, researchers must make the most of what is publicly available and important information regarding age, gender, socioeconomic status, and lifestyle choices should be inferred from whatever content online users have shared.

In this study we investigated lung cancer risk as a function of demographics and socioeconomic status leveraging openly available big online data sources. Lung cancer is the leading cause of cancer death in the United States [2] both for males and females. Because

of its high incidence, lung cancer patient and survivor stories are abundant online. It is reported that the incidence of lung cancer increases with age [3] and the most important risk factor for lung cancer is smoking tobacco [4]. In the United States, 20.5% of adult males and 15.3% of adult females are smokers [5]. Tobacco consumption is also strongly associated with socioeconomic status [6]. People with lower income and lower education have higher prevalence of smoking, which suggests an association between socioeconomic status and lung cancer risk [7]. Indeed, an observational case-control study performed by Mao et al with over 3,200 male subjects and over 5,000 female subjects [8] reported that people of low income have significantly higher lung cancer risk than people of higher income. An association between lower education and higher lung cancer risk was also reported in the Mao study.

In this paper we aimed to explore if the same association between sociodemographic factors and lung cancer risk can be captured reliably using cyber-informatics. We specifically focus on age, gender, and income status. We use a previously reported method to collect the case and control groups using online sources. Tailored natural language processing techniques are applied to extract biographical information and to classify occupational categories from employment titles for inferring income levels (Section II). Logistic regression lung cancer risk models are developed in Section III. Section IV reports cancer risk insights found using this approach as well it discusses future improvements.

II. Methods

Our approach replicates a case-control study design. We follow a two-step process to collect subjects for the case and control groups. First, we identify lung cancer patient and survivor stories available on the Internet and find their matched LinkedIn profiles. Second, we use other LinkedIn profiles with no reference to cancer as examples of control subjects. Finally, we use the information provided in the LinkedIn profiles to infer the subjects' sociodemographic information such as age and gender as well as their income level from their occupation. The extracted attributes (age, gender, income) are used to build lung cancer risk models.

A. Data Source

LinkedIn is a business-oriented social network service. It allows subscribers to create their professional profiles, including job history, work experience and education, as well as establish their online social network representing real-world professional relationships. 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 their limitations, we developed an additional step to identify LinkedIn profiles of lung cancer survivors.

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 an advanced web crawler [9]. From the life stories

of those candidate cancer subjects, we identified the subjects' names and other information using the Stanford CoreNLP [10] that helped us search and match them with profiles available in LinkedIn. Matched profiles constituted the Case Group. General LinkedIn profiles were collected to comprise the Control Group. Since there is no gold standard available to confirm that these profiles were indeed from lung cancer-free subjects, the control group must be considered as a random sample of the general population rather than the lung cancer – free population. However, at large scales such as in our study, case-control studies and case-population studies lead to the same findings.

B. Inferring Age and Gender

Since LinkedIn profiles have no direct indication of the subjects' gender and age, we developed tailored algorithms to infer such information. Gender was inferred from the first name of the LinkedIn profile utilizing the genderize.io API [11]. 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 with confidence were excluded from further analysis.

C. Inferring Income Level

We used income level as a surrogate measure of the subjects' socioeconomic status. Income level was inferred by the median earning of the subjects' present occupation. We categorized occupations based on the Occupation Classification Codes (OCC) provided by the US Census Bureau [12]. The 2010 Standard Occupation Classification Manual contains 509 sub-categories arranged into 23 major categories. We classified occupations of subjects into these 23 major OCC categories and retrieved their earnings from the median earnings table available at [13].

First, occupations of subjects were determined by the job title on the subjects' LinkedIn profiles. Then, we classified job titles to the appropriate OCC job categories. Classification was done using the marginal probability of association of words in the job titles to the OCC categories. Probabilities of words associated with the job categories in OCC were calculated by the frequency of the words appearing in the US Census 2010 Occupation Index and Industry Index [14], which contains 21,000 industry and 31,000 occupation titles. Table I lists occupations of categories and median earnings by gender.

In this study, we composed a dictionary of 6,947 words from the industry and occupation titles along with frequency of words in each occupation category, which is a probability of association of words to the job categories. For example, job title "B777 Pilot" consists of two words "B777" and "Pilot", while "B777" is not listed in the dictionary, the word "Pilot" is in the dictionary and has strong association with the job code 22, "Transportation and material moving" occupation. Thus we classify the subject with job title "B777 Pilot" in occupation code 22.

For multiple words we calculated the marginal probability of words with the association of occupation categories. For example, job title "Computer Repair Technician" consists of three words. The word "Computer" has strong association with multiple job categories, in which the strongest is code 3. However, the marginal probability of three words, "Computer",

“Repaid”, and “Technician” is strongly associated with the job code 20, “Installation, maintenance, and repair” occupation.

Note that the US Census 2010 Occupation Index and Industry Index contains no titles associated with the code 23, “Armed Forces” occupation, nor available median earnings of military job category. For the study, we added titles of ranks of US Army, Navy, and Air Force and imputed average income of all other categories as their income. We believe that this imputation rule does not affect the overall result of the study since there are only a few subjects who were employed in the armed forces.

III. Experimental Results

We retrieved 2,353 LinkedIn profiles of lung cancer survivors; 1,011 females and 1,342 males. We also randomly selected LinkedIn profiles to comprise the control group, 10,466 female subjects and 9,440 male subjects. The numbers of subjects by age group and gender are listed in Table II. Occupations of subjects were identified by the method described in Section IIc and they are listed in Table III. For subjects in the case group, we identified the job titles in the year the subjects were diagnosed with lung cancer. For control subjects, we used the latest or most recent job titles to classify occupation categories. Then, the estimated income was assigned based upon the median earnings by identified job code and gender listed in Table II.

The average earning for the case subjects was \$58,431.17, and for the control subjects was \$57,547.92. Average estimated earning for female subjects was \$49,505.09, and for male subjects was \$66,301.93. The difference of average earnings between the case and control groups was marginal, however there was notable difference between the earnings of female and male subjects illustrating the well known income gap between genders.

Age, gender, and income level of subjects were included in a logistic regression model to control the confounding effect of sociodemographic factors. The function glm in R version 3.1.0 was used in this study. Results are shown in Table IV. There are some distinct findings.

- There exists a significant increasing trend of lung cancer risk from low age to higher age, which is consistent with what is known from traditional epidemiological studies.
- Higher lung cancer risk was observed for male subjects than female subjects but the difference was not significant. This is also consistent with reported trends. The lung cancer rate has been dropping among men over the past 2 decades and closing the gap between men and women.
- Subjects with medium income have lower risk than low income group, however the difference is not statistically significant. The lung cancer risk for the higher income group is significantly lower than for the lower income subjects, which is consistent with other reports [7][8]. Please note that income status was categorized into three classes; low (<\$50,000), medium (\$50,000~\$60,000), and high (>\$60,000), with 6,714 subjects in

low income, 7,974 subjects in medium, and 7,571 subjects earned high income.

IV. Conclusion and Discussion

In this paper, we presented natural language processing methods for inferring biographical information as well as socioeconomic status of subjects from openly available LinkedIn profiles. This critical step enabled us to develop lung cancer risk models by adjusting accordingly for possible confounding factors such as age, gender, and income level. The odds ratio obtained from logistic regression clearly showed strong association between age and income level and lung cancer risk, which is consistent to other studies and reports.

There is an important limitation in the study, namely the lack of adjustment due to the subjects' smoking history. Such information is not available in the LinkedIn profiles and it is beyond the scope of this investigation.

In addition, there is a limitation with the occupation classification step. It turns out that many job titles contained the word "Manager" which led to their classification as code 1, under "Management" occupation. Therefore, all job titles which included the word "manager" were classified into the same category ignoring potentially important subcategorizations which could estimate income level more accurately.

Still, this in silico case-control study produced findings very similar to those reported in traditional cancer epidemiology. This study provided additional evidence of how big, openly available datasets can be leveraged in new and creative ways to gain valuable insights in a dynamic and cost-effective way.

Acknowledgment

This study was supported by Grant No. 1R01-CA170508-03 from the National Cancer Institute.

References

- [1]. Yoon, H.; Xu, S.; Tourassi, G. A cost-effective, case-control study on the association between breast cancer and pregnancy through web mining. Biomedical Sciences and Engineering Conference (BSEC); 2013; IEEE; 2013. p. 1-4.
- [2]. American Cancer Society. Cancer Facts & Figures 2014. American Cancer Society; Atlanta: 2014.
- [3]. Stewart, BW.; Kleihues, P., editors. World Cancer Report. Vol. Vol. 57. IARC press; Lyon: 2003.
- [4]. Zang EA, Wynder EL. Differences in lung cancer risk between men and women: examination of the evidence. Journal of the National Cancer Institute. 1996; 88(no. 3-4):183–192. [PubMed: 8632492]
- [5]. [accessed May 21, 2015] CDC – Fact Sheet – Current Cigarette Smoking Among Adults in the United States. http://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/
- [6]. CDC. Cigarette Smoking Among Adults – United States, 2001, MMWR 2008. 2008; 57(no. 45): 1221–1226.
- [7]. Devesa SS, Diamond EL. Socioeconomic and racial differences in lung cancer incidence. American Journal of Epidemiology. 1983; 118(no. 6):818–831. [PubMed: 6650483]
- [8]. Mao Y, Jinfu H, Ugnat A, Semenciw R, Fincham S. Socioeconomic status and lung cancer risk in Canada. International journal of epidemiology. 2001; 30(no. 4):809–817. [PubMed: 11511609]

- [9]. Xu S, Yoon HJ, Tourassi G. A user-oriented web crawler for selectively acquiring online content in e-health research. *Bioinformatics*. 2014; 30(no. 1):104–114. [PubMed: 24078710]
- [10]. Manning, CD.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, SJ.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; 2014; p. 55-60.
- [11]. [accessed May 21, 2015] Determine the Gender of a First Name. <http://genderize.io/#overview>
- [12]. [accessed May 20, 2015] Current Population Survey. <http://www.census.gov/cps/methodology/ioclassification.html>
- [13]. [accessed May 20, 2015] Industry and Occupation – Table Packages – People and Households. https://www.census.gov/people/io/publications/table_packages.html
- [14]. [accessed May 20, 2015] Industry and Occupation – Indexes – People and Households. <https://www.census.gov/people/io/methodology/indexes.htm>

TABLE I

The Occupation Classification Codes, Descriptions, and Median Earnings

Code	Description	Earning (Dollars)	
		Female	Male
<i>1</i>	Management	59,964	79,836
<i>2</i>	Business and financial operations	52,958	70,670
<i>3</i>	Computer and mathematical science	69,795	80,509
<i>4</i>	Architecture and engineering	64,961	79,244
<i>5</i>	Life, physical, and social science	57,617	66,913
<i>6</i>	Community and social service	41,485	43,927
<i>7</i>	Legal	61,432	116,689
<i>8</i>	Education, training, and library	44,667	55,349
<i>9</i>	Arts, design, entertainment, sports, and media	47,086	52,617
<i>10</i>	Healthcare practitioner and technical	56,710	80,723
<i>11</i>	Healthcare service	26,743	30,838
<i>12</i>	Protective service	40,760	51,159
<i>13</i>	Food preparation and serving related	20,049	22,483
<i>14</i>	Building and grounds cleaning and maintenance	20,844	27,783
<i>15</i>	Personal care and service	22,294	30,217
<i>16</i>	Sales and related	31,747	50,259
<i>17</i>	Office and administrative support	33,637	38,713
<i>18</i>	Farming, fishing, and forestry	18,998	26,271
<i>19</i>	Construction and extraction	33,236	40,078
<i>20</i>	Installation, maintenance, and repair	40,347	43,781
<i>21</i>	Production	26,544	39,170
<i>22</i>	Transportation and material moving	26,862	35,819
<i>23</i>	Armed Forces	N/A	N/A

TABLE II

Age and Gender Distribution of Case and Control Groups of the Study

Age	Cases		Controls	
	Female	Male	Female	Male
20~29	317	331	4,245	2,839
30~39	304	328	3,145	2,766
40~49	207	272	1,684	1,816
50~59	125	246	963	1,147
60+	58	165	429	872
Total	1,011	1,342	10,466	9,440

TABLE III

Number of Subjects by Occupation Code, Disease Group, and Gender

Code	Cases		Controls	
	Female	Male	Female	Male
<i>1</i>	328	505	3,851	3,537
<i>2</i>	72	91	1,172	931
<i>3</i>	20	45	178	350
<i>4</i>	13	54	312	697
<i>5</i>	29	34	161	130
<i>6</i>	34	29	327	210
<i>7</i>	7	18	184	144
<i>8</i>	63	47	437	261
<i>9</i>	79	61	491	269
<i>10</i>	91	81	551	239
<i>11</i>	41	41	534	276
<i>12</i>	18	45	135	261
<i>13</i>	7	7	51	25
<i>14</i>	0	0	1	2
<i>15</i>	10	8	78	45
<i>16</i>	29	69	541	706
<i>17</i>	58	65	604	362
<i>18</i>	4	1	12	15
<i>19</i>	17	20	150	164
<i>20</i>	13	16	102	108
<i>21</i>	72	92	522	589
<i>22</i>	6	8	68	110
<i>23</i>	0	0	4	9

TABLE IV

Odds Ratios and Corresponding 95% Confidence Intervals of the Biographical and Socioeconomical Factors

Factors		Odds Ratio
<i>Age</i>	20~29	1.00 (Reference)
	30~39	1.28 (0.99-1.64)
	40~49	1.87 (1.41-2.44)
	50~59	1.78 (1.26-2.49)
	60+	1.97 (1.30-2.90)
<i>Gender</i>	Female	1.00 (Reference)
	Male	1.31 (0.96-1.77)
<i>Earning</i>	< \$50,000	1.00 (Reference)
	\$50,000 ~ \$60,000	0.84 (0.69-1.02)
	> \$60,000	0.62 (0.38-0.97)