Journal of Human Performance in Extreme Environments

Volume 14 | Issue 1 Article 2

Published online: 1-24-2018

Investigating Sociodemographic Disparities in Cancer Risk Using Web-Based Informatics

Hong-Jun Yoon
Biomedical Sciences, Engineering, and Computing Group Health Data Sciences Institute, Oak Ridge National Laboratory, yoonh@ornl.gov

Georgia Tourassi
Biomedical Sciences, Engineering, and Computing Group Health Data Sciences Institute, Oak Ridge National Laboratory, tourassig@ornl.gov

Investigating Sociodemographic Disparities in Cancer Risk Using Web-Based Informatics

Cover Page Footnote

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This study was supported by Grant No. 1R01-CA170508-04 from the National Cancer Institute.

Investigating Sociodemographic Disparities in Cancer Risk Using Web-Based Informatics

Hong-Jun Yoon and Georgia Tourassi

Biomedical Sciences, Engineering, and Computing Group, Health Data Sciences Institute, Oak Ridge National Laboratory

Abstract

Cancer health disparities due to demographic and socioeconomic factors are an area of great interest in the epidemiological community. Adjusting for such factors is important when developing cancer risk models. However, for digital epidemiology studies relying on online sources such information is not readily available. This paper presents a novel method for extracting demographic and socioeconomic information from openly available online obituaries. The method relies on tailored language processing rules and a probabilistic scheme to map subjects' occupation history to the occupation classification codes and related earnings provided by the U.S. Census Bureau. Using this information, a case-control study is executed fully in silico to investigate how age, gender, parity, and income level impact breast and lung cancer risk. Based on 48,368 online obituaries (4,643 for breast cancer, 6,274 for lung cancer, and 37,451 cancer-free) collected automatically and a generalized cancer risk model, our study shows strong association between age, parity, and socioeconomic status and cancer risk. Although for breast cancer the observed trends are very consistent with traditional epidemiological studies, some inconsistency is observed for lung cancer with respect to socioeconomic status.

Keywords: digital epidemiology, natural language processing, case-control study, generalized linear model, obituary, cancer mortality, breast cancer, lung cancer

Introduction

Understanding and overcoming disparities in the burden of cancer is one of the overarching goals of the Cancer Moonshot initiative announced by former President Obama (Lowy & Collins, 2016). The initiative aims to galvanize research efforts and accelerate advances in cancer prevention, screening, diagnosis, and therapy. Cancer disparities are well documented across racial, ethnic, and socioeconomic groups (Braun et al., 2015; Krieger, 2005; Siegel, Miller, & Jemal, 2014; Ward et al., 2004). Studies suggest that socioeconomic factors such as income level and education are at least as important if not more so than biological factors impacting both cancer incidence and mortality rates (e.g., Danforth, 2013; Jacobs et al., 2012; Khawja et al., 2015). Income level and education are closely associated with well-known cancer risk factors such as tobacco use, poor nutrition, physical inactivity, and obesity. Poor communities and minorities have limited access to high-quality, affordable foods and fewer opportunities for safe recreational physical activity. In addition, cultural factors and professional activities influence health behaviors and attitudes, thus modifying people's cancer risk profiles. Understanding the broad spectrum and interplay of socioeconomic determinants of cancer health beyond the individual's genomic profile will enable better utilization of existing healthcare resources in the short run and faster discovery of new precision interventions in the long run.

The effect of socioeconomic factors on cancer risk is typically studied with observational, case-control studies, a well-established approach in cancer epidemiology serving as the gold standard. Traditional case-control studies require tremendous effort to recruit sufficient number of participants and to collect data from them across a range of factors. Recently, web mining has emerged as a highly promising way to leverage openly available online sources for in silico epidemiological discoveries (Khoury et al., 2013; Lam, Spitz, Schully, & Khoury, 2013). This new approach for epidemiological discovery is known as digital epidemiology. For example, in two recently published studies we demonstrated how online obituaries can be automatically collected and mined to understand the association between parity and cancer risk (Tourassi, Yoon, Xu, & Han, 2015) as well as to reliably capture spatiotemporal cancer mortality trends consistent with those reported by the national cancer surveillance program (Tourassi, Yoon, & Xu, 2016). In a different study, we demonstrated how we can leverage openly available online data to investigate the possible association between residential mobility and lung cancer risk (Yoon, Tourassi, & Xu, 2015).

http://dx.doi.org/10.7771/2327-2937.1087

Although fully automated and cost-effective, digital epidemiology has limitations due to the variable amount and granularity of information available in the open World Wide Web. For example, information related to age, gender, and socioeconomic status is not always readily available, yet all three are important confounding factors due to biases they may introduce skewing study findings. To mitigate the bias risk, traditional epidemiological studies utilize detailed questionnaires to collect information about every potential confounding factor. With open online sources, though, researchers must make the most of what is available and important information regarding age, gender, socioeconomic status, and lifestyle choices should be inferred as needed from whatever online content is available.

In this study we focus on socioeconomic disparities for selected cancer sites, namely lung and breast. Together, these two cancer types comprise 28% of new cases and 32% of cancer deaths anticipated in the United States in 2016 (Siegel, Miller, & Jemal, 2016). Specifically, lung cancer is the leading cause of cancer deaths in the United States (Siegel et al., 2016) for both males and females. It is reported that the incidence of lung cancer increases with age (Stewart & Kleihues, 2003) and the most important risk factor for lung cancer is smoking tobacco (Zang & Wynder, 1996). In the United States, 20.5% of adult males and 15.3% of adult females are smokers (CDC, 2015). Tobacco consumption is also strongly associated with socioeconomic status (CDC, 2008). People with lower income and lower education have higher prevalence of tobacco use, which suggests an association between socioeconomic status and lung cancer risk (Devesa & Diamond, 1983). Indeed, an observational case-control study performed by Mao, Hu, Ugnat, Seminciw, and Fincham (2001) with over 3,200 male subjects and over 5,000 female subjects 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 same study. Breast cancer, on the other hand, is the most frequently diagnosed cancer in females both in the United States and in Europe (Siegel et al., 2016; Stewart & Kleihues, 2003). Both demographics (e.g., ethnicity and age) as well as lifestyle factors (e.g., alcohol intake, smoking, diet, physical activity) have been found to influence breast cancer risk and to some extent prognosis and mortality (BreastCenter.org, 2016). In contrast to lung cancer, socioeconomic status has been associated with higher breast cancer risk but the association with mortality is less clear (Bouchardy, Verkooijen, & Fioretta, 2006; Ferlay et al., 2013; Lundqvist, Andersson, Ahlberg, Nilbert, & Gerdtham, 2016). For example, a study inferring socioeconomic status based on residential area did not show significant correlation between breast cancer mortality and residential area (Akinyemiju et al., 2015). The latest meta-analysis study concluded that the effects of comorbidities and lifestyle factors are difficult to delineate in breast cancer (Lundqvist et al., 2016).

The purpose of this study was to explore if the same association between socioeconomic and demographic factors and cancer risk can be captured reliably using web-based informatics. We focus specifically on breast and lung cancer, the two most prevalent cancer types (Siegel et al., 2016) for which patient and survivor stories are abundant online. Furthermore, we focus on age, gender, and income status. We use a previously reported web crawling method (Xu, Yoon, & Tourassi, 2013) to automatically derive case and control subject cohorts based on online sources. Then, we apply tailored natural language processing techniques to extract biographical information. In the following sections, a novel algorithm is developed for inferring economic status using employment information available for the collected subjects, and logistic regression cancer risk models are developed and presented. Finally, we discuss cancer risk insights derived using this novel web mining approach and compare these findings to those established with traditional cancer epidemiological studies.

Methods

To study the association between socioeconomic status and cancer risk using online content, we replicate a case-control study design. We follow a stepwise process to collect subjects for the case (cancer) and control groups. First, we automatically collect openly available obituaries and death announcements from digital sources such as the websites of U.S. newspapers, funeral homes, and social media sites. Second, we perform text mining to infer the deceased subjects' demographic information such as age and gender which are typically provided in the obituary text. Third, tailored language processing rules are applied to infer if the subjects' cause of death was breast or lung cancer as well as their occupation. The extracted attributes (age, gender, occupation) are used to build cancer risk models.

Data Sources

Online obituaries are widely available in newspaper sites and funeral homes' web pages. Although obituaries are written in various styles that can range from very formal to informal text, they have a largely similar format. Typical online obituaries include four sections; death announcement, biographical information, survivor information, and information about funeral arrangements. Intrinsically, obituaries include basic information of the deceased that is essential for our study: age (and/or birth date), cause of death, residence, and major life events (e.g., schools attended, military service, employments, organizational memberships).

We mined openly available obituary announcements from online sources. To perform automated collection of obituary contents, we used an advanced web crawler we

developed previously (Xu et al., 2013). The crawler employs an intelligent, self-adaptive mechanism to search the broad Internet for relevant content. Relevant content is available in local and national U.S. newspaper websites, homepages of U.S. funeral homes and mortuaries, and social media sites such as cancer survivor network. For this study, we focused on obituaries including the keywords "breast cancer" and "lung cancer." The crawling process was initialized with the search results of a seed query executed using a third-party commercial search engine.

We implemented an autonomous utility score estimator to assess the relevance of the crawled webpages and the embedded URLs as described in Xu et al. (2013). The utility score estimator is a supervised machine learning method, already trained with manually selected positive and negative training samples. Positive examples included breast cancer- and lung cancer-related obituaries while negative examples included unrelated webpages. Then, we applied a second verification step to select webpages representing full-length obituaries. This step also employed a supervised classification algorithm to remove irrelevant content such as obituary index pages or obituary snippets. More details about the web crawling process can be found elsewhere (Tourassi et al., 2015; Xu et al., 2013).

The Oak Ridge Site-Wide Institutional Review Board (IRB) performed expedited review and deemed the study exempt.

Sociodemographic Information Extraction

The next step was text mining of obituary contents to extract information about the deceased subjects. First, we applied the Stanford Natural Language Processing library (Manning et al., 2014) to analyze the text of the collected obituaries and select those for which the gender, parity, age at death, cause of death, and employment/occupation could be inferred. Text parsing was executed on the Titan supercomputer of the Oak Ridge Leadership Computing Facility.

The tailored rules developed to infer age, gender, parity, and cause of death were described in detail in a previous publication (Tourassi et al., 2015). Briefly, gender was inferred by counting the prevalence of male and female pronouns present in the obituary (e.g., "She passed away at her residence…"). Age of the deceased either is clearly stated in the text or was inferred based on the dates of birth and death which are often stated in the obituary text. Parity for female subjects was inferred by the listing of surviving children mentioned in the obituary or by searching for expressions such as "She was a mother of…." Surviving children were counted as biological offspring unless stated otherwise. If an obituary did not include such statements, the female subject was considered nulliparous. History of lung or breast cancer was inferred from explicit statements that the specific type of cancer was the cause of death (e.g., "He passed away after a courageous battle with lung cancer…"). Heuristic rules were applied to filter out those obituaries that may contribute to false counts. For example, simply the mention of "lung cancer" was not sufficient to consider the diseased as a member of the "case" group because there may be sentences stating the family prefers monetary contributions to a cancer research foundation rather than flowers (e.g., "In lieu of flowers, please consider donations to cancer research."). Obituaries including such sentences were not considered in this study.

For this study, we developed an additional set of heuristic rules to infer the deceased subject's occupation. Specifically, employment history was derived from explicit statements (e.g., "He worked for 34 years at the Ford assembly plant."). Since there are various types of such statements which cannot be described by the rule-based approach, we employed a supervised classification algorithm to identify sentences of employment history. Specifically, we trained a logistic regression classifier with 73 sentences describing employment history of deceased as positive examples, and 700 sentences randomly chosen from obituary text as negative examples. Based on cross-validation, the F1-score of the classifier was found to be 0.918 (precision 0.894, recall 0.943), which was considered sufficiently reliable.

Income Level Inference

We used income level as a surrogate measure of the subjects' socioeconomic status. Income level was inferred by the median earning of the deceased's occupation history. We categorized occupations based on the Occupation Classification Codes (OCC) provided by the U.S. Census Bureau (2015a). The 2010 Standard Occupation Classification Manual contains 509 subcategories arranged into 23 major categories (Table 1). We classified the subjects' occupations into these 23 major OCC categories and retrieved their earnings from the median earnings table (U.S. Census Bureau, 2015b).

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 U.S. Census 2010 Occupation Index and Industry Index (U.S. Census Bureau, 2015c), which contains 21,000 industry and 31,000 occupation titles. Table 1 lists the occupation categories and median earnings by gender. We identified the estimated earnings from the occupation history statements of subjects. Obituaries with no mention of occupation history were excluded from the 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 multiple words we calculated the marginal probability of words with the association of occupation categories.

Note that the U.S. Census 2010 Occupation Index and Industry Index contains no titles associated with code 23, "armed forces" occupation, nor available median earnings

Table 1
The occupation classification codes; high-level aggregation to six groups, code descriptions, and median earnings.

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

for the 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. A negligible number of obituaries included subjects with occupation history in armed forces.

Statistical Analysis

Using a case-control study design, age-adjusted odds ratios and 95% confidence intervals were calculated by the generalized linear model (GLM), which allows controlling for potential confounders. In this study, gender, parity, and income level of subjects were applied to the GLM implemented in R version 3.3.1 with a binomial family.

Results

We retrieved 4,643 obituaries of female breast cancer death and 6,274 obituaries of lung cancer death; 2,289 females and 3,985 males. The number of obituaries of non-cancer deaths collected to comprise the control group was 37,451; 15,870 females and 21,581 males. Note that the collected subjects included in the study are those for whom all necessary information pieces (i.e., gender, age, parity, cause of death, occupation) could be inferred with high confidence. Table 2 shows the number and age distribution of the cancer cases and controls. The distribution and average age of subjects according to the number of offspring are shown in Table 3.

The numbers of subjects by occupation are listed in Table 4. Estimated average earning for female breast cancer subjects was $43,957.89, for female lung cancer subjects average annual income was $40,764.16, and for male lung cancer subjects $51,889.37. For the control group, the estimated average annual earnings were $38,628.69 for females and $50,305.76 for males. Overall, the average earning for female subjects was $39,707.37 and for male subjects was $50,552.59. The difference of average earnings between the case and control groups was marginal; however, there was a notable difference between the earnings of female and male subjects illustrating the well-known income gap between genders.

Table 5 shows the risk for breast and lung cancer with respect to estimated income level based on a logistic regression model, adjusting for confounding factors such as

Table 2
Number and age distribution of cancer cases and controls of the study.

Age	Breast	Lung		Controls	
	Female (N = 4,643)	Female (N = 2,289)	Male (N = 3,985)	Female (N = 15,870)	Male (N = 21,581)
20–29	1.87%	0.74%	1.10%	1.30%	1.28%
30–39	6.01%	1.27%	1.83%	1.41%	1.39%
40–49	14.49%	5.16%	5.45%	3.39%	3.38%
50–59	23.39%	15.55%	15.06%	6.71%	6.30%
60–69	20.89%	25.08%	26.50%	11.06%	11.59%
70–79	15.94%	30.32%	28.71%	18.70%	18.28%
80+	17.40%	21.89%	21.36%	57.42%	57.78%

Table 3
Average age at death of cancer cases and controls by parity.

		Null	1–2	3–4	5+
Breast	**No. cases**	15.44%	55.24%	19.15%	10.17%
	Age (σ)	54.85 (16.37)	60.15 (16.43)	62.70 (15.89)	65.36 (14.01)
Lung female	**No. cases**	15.99%	52.12%	19.66%	12.23%
	Age (σ)	64.84 (13.81)	67.49 (12.89)	66.79 (12.71)	68.79 (12.11)
Lung male	**No. cases**	15.23%	51.59%	20.90%	12.27%
	Age (σ)	64.04 (13.92)	66.16 (13.44)	66.40 (13.68)	70.06 (11.07)
Controls female	**No. cases**	13.29%	53.55%	19.92%	13.25%
	Age (σ)	73.86 (17.74)	76.16 (15.48)	75.76 (14.75)	77.35 (13.08)
Controls male	**No. cases**	13.71%	52.85%	20.29%	13.16%
	Age (σ)	73.89 (17.71)	76.41 (15.46)	75.97 (14.36)	76.94 (13.20)

Table 4
Occupation distribution of cancer cases and control groups by occupation code.

Code	Description	Breast	Lung		Controls	
			Female	Male	Female	Male
1	Management	13.90%	12.68%	15.44%	8.70%	15.05%
2	Business and financial operations	1.63%	1.32%	0.58%	1.23%	0.63%
3	Computer and mathematical science	1.87%	2.64%	1.75%	1.13%	1.47%
4	Architecture and engineering	1.13%	0.40%	4.85%	0.41%	3.22%
5	Life, physical, and social science	0.35%	0.26%	0.58%	0.31%	0.21%
6	Community and social service	1.42%	1.32%	0.39%	0.72%	0.56%
7	Legal	0.01%	0.00%	0.00%	0.00%	0.00%
8	Education, training, and library	10.94%	9.91%	1.75%	9.93%	2.73%
9	Arts, design, entertainment, sports, media	1.42%	1.98%	0.97%	0.61%	0.49%
10	Healthcare practitioner and technical	14.86%	13.34%	4.85%	13.41%	3.08%
11	Healthcare support	4.10%	3.96%	0.97%	2.87%	0.91%
12	Protective service	2.31%	2.11%	6.41%	0.72%	5.60%
13	Food preparation and serving related	0.83%	0.92%	0.68%	3.07%	0.70%
14	Building and grounds cleaning	0.00%	0.00%	0.00%	0.00%	0.00%
15	Personal care and service	0.50%	0.40%	0.19%	0.61%	0.21%
16	Sales and related	3.29%	5.42%	3.59%	2.76%	3.36%
17	Office and administrative support	13.81%	14.27%	9.90%	19.04%	9.17%
18	Farming, fishing, and forestry	0.14%	0.26%	0.00%	0.20%	0.21%
19	Construction and extraction	1.45%	0.40%	8.35%	0.72%	7.49%
20	Installation, maintenance, and repair	2.66%	2.51%	6.31%	2.66%	8.19%
21	Production	21.86%	24.17%	27.57%	29.27%	29.32%
22	Transportation and material moving	1.53%	1.72%	4.85%	1.64%	7.42%
23	Armed forces	0.00%	0.00%	0.00%	0.00%	0.00%

Table 5
Odds ratios and corresponding 95% confidence intervals of the biographical and socioeconomic factors.

Factors		Odds Ratio
	Breast cancer	
Parity	0	1.00 (Reference)
	1–2	0.89 (0.81–0.99)
	3–4	0.83 (0.74–0.93)
	5+	0.67 (0.58–0.76)
Earning (quartile)	1 (< $26,544)	1.00 (Reference)
	2 ($26,544–$33,637)	1.25 (1.14–1.38)
	3 ($33,637–$56,710)	2.11 (1.93–2.30)
	4 (> $56,710)	2.38 (2.14–2.65)
	Lung cancer (female)	
Parity	0	1.00 (Reference)
	1–2	0.81 (0.71–0.92)
	3–4	0.82 (0.71–0.95)
	5+	0.77 (0.65–0.91)
Earning (quartile)	1 (< $26,544)	1.00 (Reference)
	2 ($26,544–$33,637)	1.18 (1.05–1.34)
	3 ($33,637–$56,710)	1.51 (1.35–1.70)
	4 (> $56,710)	1.69 (1.47–1.95)
	Lung cancer (male)	
Parity	0	1.00 (Reference)
	1–2	0.88 (0.80–0.97)
	3–4	0.93 (0.83–1.04)
	5+	0.84 (0.74–0.96)
Earning (quartile)	1 (< $39,170)	1.00 (Reference)
	2 ($39,170–$40,078)	1.28 (1.12–1.45)
	3 ($40,078–$55,349)	1.21 (1.11–1.32)
	4 (> $55,349)	1.29 (1.18–1.40)

gender, age, and parity (for females only). The table shows that both breast and lung cancer risks decrease for females with increasing number of offspring. The protective effect of parity is well established for breast cancer with traditional epidemiological studies (BreastCenter.org, 2016; Tourassi et al., 2015). The protective effect is also clear based on Table 3 which shows that the age at death increases sharply as parity increases for females with breast cancer, while such a trend is not as notable for female controls. For lung cancer, epidemiological studies are inconclusive with meta-analysis suggesting no association between parity and risk of lung cancer (BreastCenter.org, 2016; Tourassi et al., 2015). Our study suggests a protective effect based on Tables 2 and 5; however, this is neither as pronounced nor as linearly dependent on number of offspring as with breast cancer. Since our study could not adjust for important lifestyle factors that are always included in lung cancer risk models (i.e., smoking), we should not draw any further conclusions.

With respect to socioeconomic factors, Table 5 shows that income level is associated with both breast and lung cancer risk. For female breast cancer, our study confirms findings from traditional epidemiology showing a trend of increasing breast cancer risk with increasing income level. A similar trend (albeit less pronounced) was observed for both female and male lung cancer, with the trend being stronger for females. These findings are inconsistent with traditional epidemiological studies, which suggest an inverse relationship. Such inconsistency could be attributed to lack of adjustment for smoking history, due to lack of such detailed information in obituaries.

Discussion

In the "big data" era, web mining has emerged as a powerful approach to collect very large amounts of digital data openly available from diverse sources. Developing and deploying scalable web mining tools which can leverage the supercomputing resources available at the Department of Energy are important steps to accelerate information collection and enable data-driven knowledge discovery across many different domains. Our study demonstrates how such capability could transform digital epidemiology.

Specifically, in this paper we presented a web mining method for studying demographic and socioeconomic cancer health disparities leveraging online obituaries, a non-traditional openly available data source. By applying tailored text mining methods for inferring biographical information and socioeconomic status of subjects, we were able to replicate in silico a case-control epidemiological study of the association of gender, parity, and socioeconomic status with breast and lung cancer risk.

To a large extent our findings are consistent with those reported in traditional cancer epidemiology. This is particularly true regarding the association of age, gender, and parity. However, our findings demonstrate a strong negative association between income level and cancer risk. Although a similar association has been identified with carefully designed epidemiological studies for breast cancer, the opposite has been reported for lung cancer, which raises some questions about using obituaries for deriving socioeconomic information or whether the lack of adjustment for smoking history is an important limiting factor. Since our study did not incorporate any geographical information, it is possible to adjust for smoking history by applying "correction factors" for smoking history. For example, one possible way is to leverage population level statistics of tobacco use that is available for all U.S. states and different age ranges. In a future study we will attempt to develop and apply such correction factors as well as study disparities across occupations and geographical locations.

Overall, obituaries have limited details about individuals' lifestyle choices, major life events, occupation history, and date of cancer diagnosis. Therefore, there is a higher level of ambiguity which may compromise the reliability of epidemiological discovery. Similar ambiguity and thus limitations exist with using occupation classification codes to infer individual income level. Still, this in silico case-control study provides additional evidence of how large, openly available datasets can

be leveraged in new and creative ways to gain valuable insights in a dynamic and cost-effective way. Although web-based informatics approaches are not ready to replace carefully designed epidemiological studies, they are a very promising complementing technology for knowledge discovery, hypotheses generation, and additional validation.

In summary, the study we presented in this paper exemplifies the role of big data, supercomputing, and human–computer interaction to enable biomedical knowledge discovery leveraging non-traditional data sources. The presented web-mining technology and general approach are extensible to other application domains for which "user profiling" is important, namely business, marketing, and surveillance.

Acknowledgments

This study was supported by Grant No. 1R01-CA170508-04 from the National Cancer Institute. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S., Department of Energy under Contract No. DE-AC05-00OR22725.

References

Akinyemiju, T. F., Genkinger, J. M., Farhat, M., Wilson, A., Gary-Webb, T. L., & Tehranifar, P. (2015). Residential environment and breast cancer incidence and mortality: A systematic review and meta-analysis. *BMC Cancer*, *15*(1), 191.

Bouchardy, C., Verkooijen, H. M., & Fioretta, G. (2006). Social class is an important and independent prognostic factor of breast cancer mortality. *International Journal of Cancer*, *119*(5), 1145–1151.

Braun, K. L., Stewart, S., Baquet, C., Berry-Bobovski, L., Blumenthal, D., Brandt, H. M.,…Espinoza, P. (2015). The National Cancer Institute's Community Networks Program Initiative to reduce cancer health disparities: Outcomes and lessons learned. *Progress in Community Health Partnerships: Research, Education, and Action*, *9*, 21.

BreastCenter.org. (2016, July 21). Retrieved from http://www.breastcancer.org

Centers for Disease Control and Prevention. (2008). Cigarette smoking among adults—United States, 2007. MMWR. *Morbidity and Mortality Weekly Report*, *57*(45), 1221.

Centers for Disease Control and Prevention. (2015, May 21). Fact sheet—Current cigarette smoking among adults in the United States. Retrieved from http://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/

Danforth Jr, D. N. (2013). Disparities in breast cancer outcomes between Caucasian and African American women: A model for describing the relationship of biological and nonbiological factors. *Breast Cancer Research*, *15*(3), 208.

Devesa, S. S., & Diamond, E. L. (1983). Socioeconomic and racial differences in lung cancer incidence. *American Journal of Epidemiology*, *118*(6), 818–831.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W. W., Comber, H.,…Bray, F. (2013). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, *49*(6), 1374–1403.

Jacobs, B. L., Montgomery, J. S., Zhang, Y., Skolarus, T. A., Weizer, A. Z., & Hollenbeck, B. K. (2012). Disparities in bladder cancer. *Urologic Oncology: Seminars and Original Investigations*, *30*(1), 81–88.

Khawja, S. N., Mohammed, S., Silberfein, E. J., Musher, B. L., Fisher, W. E., & George Van Buren, I. I. (2015). Pancreatic cancer disparities in African Americans. *Pancreas*, *44*(4), 522–527.

Khoury, M. J., Lam, T. K., Ioannidis, J. P., Hartge, P., Spitz, M. R., Buring, J. E.,…Herceg, Z. (2013). Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology, Biomarkers & Prevention*, *22*(4), 508–516.

Krieger, N. (2005). Defining and investigating social disparities in cancer: Critical issues. *Cancer Causes and Control*, *16*(1), 5–14.

Lam, T. K., Spitz, M., Schully, S. D., & Khoury, M. J. (2013). "Drivers" of translational cancer epidemiology in the 21st century: Needs and opportunities. *Cancer Epidemiology, Biomarkers & Prevention*, *22*(2), 181–188.

Lowy, D. R., & Collins, F. S. (2016). Aiming high—Changing the trajectory for cancer. *New England Journal of Medicine*, *374*(20), 1901–1904.

Lundqvist, A., Andersson, E., Ahlberg, I., Nilbert, M., & Gerdtham, U. (2016). Socioeconomic inequalities in breast cancer incidence and mortality in Europe—A systematic review and meta-analysis. *European Journal of Public Health*, *26*(5), 804–813.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations* (pp. 55–60). Baltimore, MD: ACL.

Mao, Y., Hu, J., Ugnat, A. M., Semenciw, R., & Fincham, S. (2001). Socioeconomic status and lung cancer risk in Canada. *International Journal of Epidemiology*, *30*(4), 809–817.

Siegel, R. L., Miller, K. D., & Jemal, A. (2014). Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, *64*(1), 9–29.

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, *66*(1), 7–30.

Stewart, B. W., & Kleihues, P. (Eds.). (2003). *World cancer report* (Vol. 57). Lyon, France: IARC Press.

Tourassi, G., Yoon, H. J., & Xu, S. (2016). A novel web informatics approach for automated surveillance of cancer mortality trends. *Journal of Biomedical Informatics*, *61*, 110–118.

Tourassi, G., Yoon, H. J., Xu, S., & Han, X. (2015). The utility of web mining for epidemiological research: Studying the association between parity and cancer risk. *Journal of the American Medical Informatics Association*, *23*(3), 588–595.

U.S. Census Bureau. (2015a, May 20). Current population survey. Retrieved from http://www.census.gov/cps/methodology/ioclassification.html

U.S. Census Bureau. (2015b, May 20). Industry and occupation—Table packages—People and households. Retrieved from https://www.census.gov/people/io/publications/table_packages.html

U.S. Census Bureau. (2015c, May 20). Industry and occupation—Indexes—People and households. Retrieved from https://www.census.gov/people/io/methodology/indexes.htm

Ward, E., Jemal, A., Cokkinides, V., Singh, G. K., Cardinez, C., Ghafoor, A., & Thun, M. (2004). Cancer disparities by race/ethnicity and socioeconomic status. *CA: A Cancer Journal for Clinicians*, *54*(2), 78–93.

Xu, S., Yoon, H. J., & Tourassi, G. (2013). A user-oriented web crawler for selectively acquiring online content in e-health research. *Bioinformatics*, *30*(1), 104–114.

Yoon, H. J., Tourassi, G., & Xu, S. (2015, March). Residential mobility and lung cancer risk: Data-driven exploration using internet sources. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 464–469). Cham, Switzerland: Springer International Publishing.

Zang, E. A., & Wynder, E. L. (1996). Differences in lung cancer risk between men and women: Examination of the evidence. *Journal of the National Cancer Institute*, *88*(3–4), 183–192.