Web Mining for Epidemiological Research: Assessing its Utility in Exploring the Association Between Parity and Cancer Risk

Georgia Tourassi[a], Hong-Jun Yoon[a], Songhua Xu[b], Xuesong Han[c]

[a] Health Data Sciences Institute, Biomedical Science and Engineering Center,

Oak Ridge National Laboratory, Oak Ridge, TN 37831

[b] Information Systems Department,

New Jersey Institute of Technology, Newark, NJ 07102

[b] Surveillance and Health Services Research,

American Cancer Society, Atlanta, GA 30303

Corresponding Author:

Georgia D. Tourassi, PhD
Oak Ridge National Laboratory
PO Box 2008 MS6085
Oak Ridge, TN 37831-6085
Phone: (865) 576-4829
E-mail: tourassig@ornl.gov

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).

ABSTRACT

Background: The World Wide Web has emerged as a powerful data source for epidemiological studies related to infectious disease surveillance. However, its potential for cancer-related epidemiological discoveries is largely unexplored.

Methods: Using advanced web crawling and tailored information extraction procedures we automatically collected and analyzed the text content of 79,394 online obituary articles published between 1998 and 2014. The collected data included 51,911 cancer (27,330 breast; 9,470 lung; 6,496 pancreatic; 6,342 ovarian; 2,273 colon) and 27,483 non-cancer cases. With the derived information, we replicated a case-control study design to investigate the association between parity and cancer risk. Age-adjusted odds ratios (ORs) with 95% confidence intervals (CIs) were calculated for each cancer type and compared to those reported in large-scale epidemiological studies.

Results: Parity was found to be associated with a significantly reduced risk of breast cancer (OR=0.78, 95% CI = 0.75 to 0.82), pancreatic cancer (OR=0.78, 95% CI = 0.72 to 0.83), colon cancer (OR=0.67, 95% CI = 0.60 to 0.74), and ovarian cancer (OR=0.58, 95% CI = 0.54 to 0.62). Marginal association was found for lung cancer prevalence (OR=0.87, 95% CI = 0.81 to 0.92). The linear trend between multi-parity and reduced cancer risk was dramatically more pronounced for breast and ovarian cancer than the other cancers included in the analysis.

Conclusion: This large web-mining study on parity and cancer risk produced findings very similar to those reported with traditional observational studies. It may be used as a

promising strategy to generate study hypotheses for guiding and prioritizing future epidemiological studies.

INTRODUCTION

Over the past decade the World Wide Web has been transforming radically the landscape of medical research and healthcare delivery (1-3). Online news sources and digital social networks have empowered a wide range of epidemiological applications such as syndromic surveillance (4-6), monitoring population health-related interests, sentiment, and behaviors (e.g., 7-15), as well as drug performance monitoring (16,17). Nevertheless, the potential and reliability of openly available online content for cancer-related epidemiological discoveries remain largely unknown. Thus far studies have focused on sentiment analysis, online support, and topics of interest for cancer patients (12,21,22). The current study evaluated the feasibility of web mining as a big data-driven knowledge discovery mechanism for exploring the association between parity and cancer risk.

To date, epidemiological studies have investigated parity as an independent risk factor for many types of women's cancers (23). Particularly with respect to breast cancer, the protective association between childbearing and breast cancer has been well established with several prospective, retrospective, and meta-analysis studies conducted in the US and abroad (24-30). For ovarian cancer, several international studies found that the risk across all histologies is inversely related to parity (31-40).

The association between parity and cancer risk has been studied for other cancers that are common between both genders. For pancreatic cancer there are a few studies with conflicting findings (41). A comprehensive meta-analysis study reported borderline statistically significantly lower pancreatic cancer risk for parous women (42). A stronger association was observed with only

the case-control studies included in meta-analysis. For lung cancer, the reported findings are highly heterogeneous and inconsistent. Meta-analysis of 16 case-control and cohort studies suggested no effect of parity on lung cancer risk (43). Similar inconsistencies were observed for colorectal cancer as well. Some studies reported a protective effect with parity (44-48) while other studies found no association (49-51). The largest meta-analysis of prospective studies found no association between colorectal cancer risk and parity (52).

In this study we adopted a cyber-informatics approach to perform in-silico large scale, case-control experiments for studying the relationship between parity and cancer risk. We used online obituaries of women as the data source. By comparing our findings with the findings reported in traditional epidemiological studies, we examined the feasibility, reliability, and limitations of web mining and cyber-informatics for epidemiological discoveries in cancer.

SUBJECTS AND METHODS

Study Population

We performed this study using online obituary announcements published in US newspapers. Such content is widely and openly available on the Internet. Obituaries typically include content regarding the deceased person's family members and cause of death, both essential for this investigation. Furthermore, the language of obituary articles is often standardized, which facilitates automated computer parsing and text analysis. Our study focused on online obituaries of females for whom cancer is the stated or inferred cause of death (e.g. *"She passed away after a courageous battle with breast cancer…"*) and obituaries of females for whom there is no mention of cancer.

These two types of obituaries constituted the case and control groups respectively. No other inclusion or exclusion criteria were applied. Institutional review board approval was obtained prior to the study.

Data Collection and Information Retrieval

The data collection process was fully automated and included: (a) a keyword-based Internet search for collecting obituaries, and (b) text parsing for identifying those obituaries which contained the necessary information elements (i.e., gender, age at death, offsprings, cause of death).

First, an advanced web crawler developed in our laboratory (53) was deployed to search the Internet for obituary announcements published in online US newspapers. The search targeted obituaries including as keywords the types of cancer considered in this study: "breast cancer", "ovarian cancer", "lung cancer", "pancreatic cancer", and "colon cancer". A similar web crawling search was deployed for collecting obituaries which did not include the keyword "cancer". The crawling process produced a total of 172,765 cancer-related obituaries (Breast: 57,293; Ovarian: 17,770; Lung: 50,429, Pancreatic: 36,280; Colon: 10,993) and 1,039,618 non-cancer related obituaries.

The body text of the collected obituaries was subsequently analyzed using tailored Natural Language Processing (NLP) procedures to automatically extract the gender, age at death, and offspring information about the deceased subjects. This step was implemented using the Stanford NLP Library (54). Several customized rules were implemented:

- Gender: To infer the gender of the deceased person gender-specific pronouns mentioned in the obituary were counted. Then, a straightforward inference rule was applied stating that the deceased person is assigned to the gender indicated by the most related pronouns counted. If the numbers of male and female pronouns counted were equal, no dominant gender could be identified and the obituary was excluded from further analysis.

- Age: In many obituaries, the age at death is explicitly stated in the first sentence along with the subject's name, e.g.: *"Jane Doe, 60, passed away…"*, *"Jane Doe, age 55, passed peacefully…"*. If no such statement was available, age was calculated by automatically detecting dates of birth and death, e.g.: *"Jane Doe passed away Sunday, January 1, 2012… She was born July 1, 1940…"*. If the there was no explicit statement, then the newspaper publication year was considered as the year of death. The age at death was calculated by simple subtraction of the birth year from the death year. Obituary articles for which the subject's age could not be detected or inferred were excluded from further analysis.

- Parity: History of childbirth was inferred by the listing of surviving offsprings mentioned in the obituary, e.g.: *"She is survived by her son(s)… and daughter(s)…", "Preceded in death by her son(s)… and daughter(s)…"*. In other cases, number of offsprings was inferred by searching for expressions such as *"She was a mother of …"*. Identified offsprings were counted as biological children unless clearly stated as stepchildren or adopted, in which case they were not included. If the obituary did not include such statements, the subject was considered nulliparous.

- Cause of Death: To infer whether cancer was the cause of death, a sequence of logical exclusion and inclusion rules was executed on those obituaries with the keyword *"<type> cancer"*.

Rule 1: Since many obituaries encourage donations to cancer organizations, all obituary sentences with phrases such as *"may be made to..." , "In lieu of flowers,...", "Donations of sympathy..."* were first removed to filter out obituaries that could contribute to false counts and simplify the inference process.

Rule 2: The remaining obituary text body was searched for explicit statements that a specific cancer type was the cause of death, e.g.: *"She passed away after a lengthy battle with breast cancer…"*. If no such statement was found, text parsing continued by applying the additional heuristic inference rule shown below.

Rule 3: Death was attributed to a "*<type>* cancer" if the obituary (i) contained the phrase "*<type>* cancer" and (ii) did not contain phrases implying that the deceased person was a cancer survivor (e.g., *'<type> cancer survivor', 'survivor of <type> cancer', 'surviving <type> cancer'*).

Cancer-related obituaries meeting none of the above conditions were excluded from further consideration. Finally, the resulting set of eligible obituaries was automatically analyzed once more to eliminate duplicate articles (e.g., obituaries of the same individual appearing in multiple different online sites). Matching was done based on the subject's name and age. The whole data collection process (web crawling and text parsing) for the cancer cases was terminated when the harvest rate for each cancer type declined significantly. Harvest rate is a commonly used performance metric for web crawlers which measures a web crawler's capability of obtaining new relevant records. In total, there were 51,911 eligible cancer subjects collected; 27,330 breast, 9,470 lung, 6,496 pancreatic, 6,342 ovarian, and 2,273 colon cancer cases. The data collection process for

the control group was interrupted after the sample size reached roughly the size of the breast cancer group, which was the largest among the cancer types included in this study.

The collected obituaries were published between 1998 and 2014. Due to the considerable computational demands of the text parsing stage, the NLP platform 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.

Statistical Analysis

Age was compared between cases and controls using Student's *t* test. Using a case-control study design, age-adjusted odds ratios (ORs) and 95% confidence intervals (CIs) were calculated using the Mantel-Haenszel procedure based upon the parity odds in the case and in the control groups. Both groups were stratified by age in 5-year intervals. The 95% CIs were estimated with bootstrapping. The primary analysis compared parous women (>=1 child) with nulliparous women. ORs and CIs were also derived for each level of parity (0, 1, 2, .., 6, >=7 children). The linear trend between increasing parity and cancer risk was also tested using the Mantel-Haenszel procedure. Statistical analysis was performed using R software, version 3.1.0 (55).

RESULTS

Figure 1 shows the distribution of the 51,911 cancer cases collected per year of online publication of the obituaries. The increasing number of collected cases from recent years reflects the increasing online presence of local US newspapers. All 50 states were represented in the dataset. Most cancer cases were collected from online obituaries published in newspapers from California, Texas, Florida, Ohio, Pennsylvania, New York, Massachusetts and Illinois. Of those states, all are the seven most populous US states with the exception of Massachusetts which ranks 14th according to the 2013 US Census data.

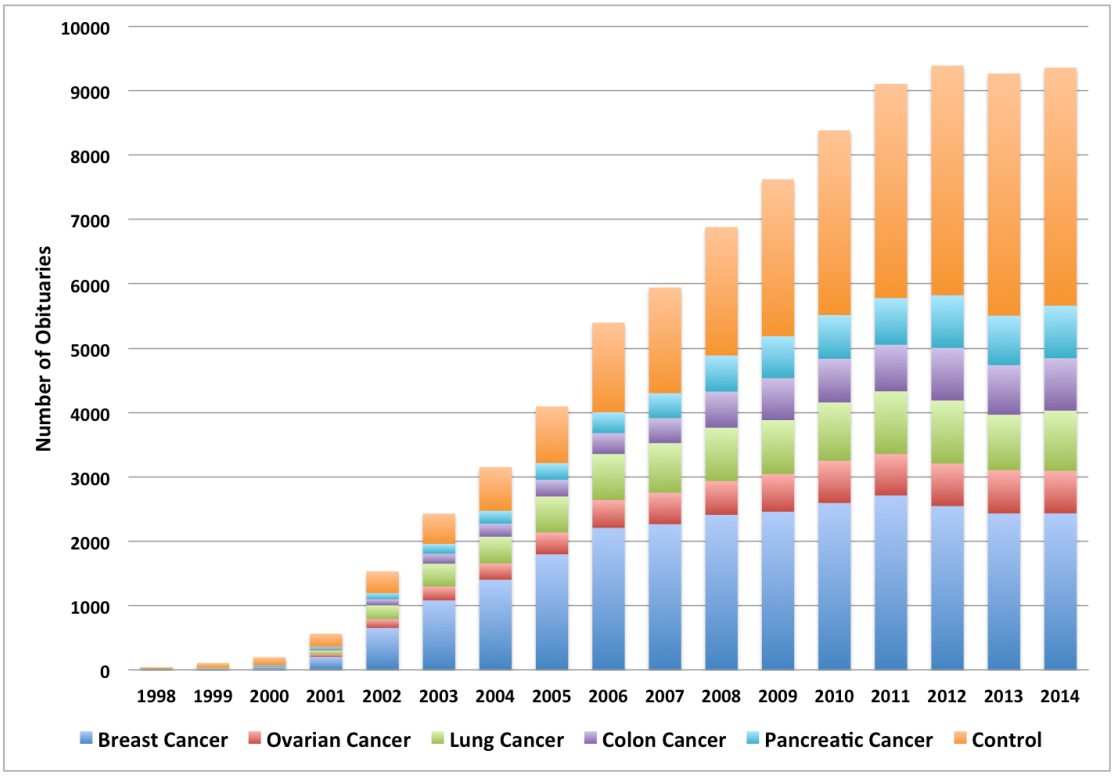

Figure 1: Number of collected obituaries included in the study by cancer type and year

Table 1 shows the number and age distribution of the cancer cases and controls collected for the study. The distribution and average age of subjects according to the number of offsprings are shown in Table 2.

Table 1: Number and age distribution of cancer cases and controls.

Age in years	Cancer Cases					Controls (N=27,483)
	Breast (N=27,330)	Ovarian (N=6,342)	Lung (N=9,470)	Colon (N=2,273)	Pancreatic (N=6,496)	
20-29	2.04%	1.37%	1.07%	2.02%	0.95%	1.28%
30-39	5.77%	2.89%	1.64%	5.28%	1.34%	1.53%
40-49	14.56%	9.04%	5.61%	11.97%	6.22%	3.45%
50-59	23.37%	24.05%	15.49%	19.31%	17.50%	6.73%
60-69	21.58%	29.63%	25.09%	20.37%	26.12%	11.29%
70-79	15.46%	21.02%	30.00%	18.79%	26.74%	18.32%
> 80	17.22%	12.02%	21.11%	22.26%	21.12%	57.39%
TOTAL	100%	100%	100%	100%	100%	100%

Table 2: Average age at death of cases and controls by parity

	Parity							
	Nulliparous		1-2		3-4		5+	
	N (%)	Age (±σ)	N (%)	Age (±σ)	N (%)	Age (±σ)	N (%)	Age (±σ)
	Cancer Cases							
Breast	16.39%	57.0 (16.5)	54.21%	62.3 (16.4)	19.13%	64.1 (16.0)	10.27%	67.5 (14.7)
Ovarian	20.64%	59.0 (13.5)	48.58%	63.8 (13.2)	18.92%	65.5 (12.7)	11.86%	68.1 (13.0)
Lung	15.25%	65.5 (14.0)	51.75%	68.3 (13.6)	20.16%	69.3 (13.1)	12.84%	71.8 (11.7)
Colon	18.13%	59.0 (17.4)	51.43%	65.2 (17.1)	17.16%	66.6 (15.2)	13.29%	68.9 (14.6)
Pancreatic	16.70%	64.7 (14.3)	50.97%	67.8 (13.5)	19.64%	69.0 (13.5)	12.68%	71.4 (11.7)
	Controls							
	13.60%	75.6 (18.1)	52.84%	78.4 (15.7)	20.52%	77.7 (14.7)	13.04%	79.1 (13.5)

The average age at death for the cancer cases was significantly lower than that for controls for all cancer types (all two-tailed p-values < 2.20e-16). For the cancer group, average age at death increased with increasing parity. No such trend was observed for the control group.

Table 3 shows the risk of cancer among parous and nulliparous women for each cancer type. While the cancer risk for parous women was significantly lower than nulliparous for all cancer

types, the protective effect of parity was much less pronounced for lung cancer than all other cancer types considered in this study.

Table 3: Age-adjusted odds ratios OR* with 95% confidence intervals (95% CIs) according to number of offsprings for each cancer type.

Parity	Breast Cancer	Ovarian Cancer	Lung Cancer	Colon Cancer	Pancreatic Cancer
Nulliparous (Reference)	1.00	1.00	1.00	1.00	1.00
Parity					
1	0.87 (0.82-0.92)	0.60 (0.55-0.65)	0.91 (0.84-0.98)	0.74 (0.65-0.84)	0.80 (0.74-0.87)
2	0.84 (0.80-0.88)	0.60 (0.56-0.65)	0.87 (0.81-0.94)	0.67 (0.59-0.76)	0.81 (0.74-0.87)
3	0.76 (0.71-0.80)	0.59 (0.54-0.65)	0.81 (0.74-0.89)	0.57 (0.48-0.67)	0.74 (0.67-0.81)
4	0.66 (0.61-0.71)	0.50 (0.44-0.56)	0.86 (0.77-0.95)	0.60 (0.49-0.72)	0.72 (0.64-0.81)
5	0.61 (0.56-0.67)	0.57 (0.49-0.66)	0.83 (0.73-0.95)	0.61 (0.48-0.78)	0.72 (0.63-0.84)
6	0.64 (0.57-0.72)	0.52 (0.44-0.63)	0.84 (0.73-0.98)	0.79 (0.61-1.02)	0.81 (0.69-0.96)
7+	0.55 (0.50-0.60)	0.52 (0.45-0.60)	0.79 (0.70-0.90)	0.68 (0.55-0.86)	0.70 (0.61-0.82)
All parous	0.78 (0.75-0.82)	0.58 (0.54-0.62)	0.87 (0.81-0.92)	0.67 (0.60-0.74)	0.78 (0.72-0.83)

The positive association between increased parity and lower cancer risk was significant for all cancer types. Although all linear trend tests were found to be statistically significant, the trend was dramatically more pronounced for breast (χ^2 = 301.60, p < 2.20e-16) and ovarian cancer (χ^2 = 121.45, p < 2.20e-16), less pronounced for pancreatic cancer (χ^2 = 38.95, p = 4.35e-10), and least pronounced for colon (χ^2 = 24.69, p = 6.75e-07) and lung cancer (χ^2 = 21.33, p = 3.87e-06).

DISCUSSION

Our study investigated the feasibility of intelligently mining a non-traditional online data source for epidemiological discoveries related to cancer. Using web mining with advanced cyber-informatics tools, we were able to collect automatically a very large number of eligible subjects and execute in silico a case-control study on the association between parity and cancer risk. The investigation focused on the five cancer types mentioned more often in the obituaries. Interestingly, these are also the cancer types associated with the highest death numbers in women in the US (56).

Our findings demonstrate that there is a significant independent association between parity and the five cancer types considered. The association is dramatically more pronounced for breast and ovarian cancer and least pronounced for lung cancer. The protective effect observed for breast, ovarian, and pancreatic cancer is very similar to that reported previously with traditional case-control observational studies. Specifically, it is reported that there is 18%-27% lower breast cancer risk for parous women (24-29). Our study showed 22% lower risk. For ovarian cancer, our study found 42% lower risk, well within the 32%-60% range reported in epidemiological studies conducted in the US (38-40). Agreement between our findings and epidemiological literature was

observed for pancreatic cancer as well. We found 22% lower pancreatic risk for parous women, compared to 28% reported in the latest meta-analysis of 11 case-control studies conducted in the US (42). For colon cancer, the 33% protective effect observed in this study is only in partial agreement with the literature since there are conflicting reports. For example, a recent US study (48) reported 20% reduced risk for postmenopausal women with no history of hormone therapy use and more than 5 pregnancies, which is similar to the age-adjusted odds ratios our study found for 5, 6, and 7+ offsprings. It should be noted that our study focused on colon cancer (since the term "colorectal cancer" was not mentioned often in the obituaries), which may have also contributed in the discrepancy. It is interesting that the only recent study we found separating colon and rectal cancers, reported approximately 40%–60% elevated risk of colon cancer for nulliparous women but no increased risk of rectal cancer (49). Finally, the 13% lower lung cancer risk for parous women was the lowest one observed in our study. The general trend is in full agreement with established knowledge that out of the five types of cancer included in this study, lung cancer is the only one for which parity appears to have marginal if any protective effect. The impact of multi-parity observed across all five cancer types was also consistent with established epidemiological knowledge. Namely, every additional pregnancy appears to reduce further breast and ovarian cancer risk while the same beneficial effect is not well established for the other three cancer types.

One potential criticism for the study is that using obituaries as a data source may introduce a selection bias towards parous women. Since the deceased person's family members typically publish obituary announcements, there is potentially a high risk of overestimating parity from

obituaries. In our study the prevalence of parity among women based on the obituaries collected for the control group was 86.4% (Table 2). The National Survey of Family Growth published by CDC/NCHS in 2012 reported that the probability of a US woman having had a birth by the age of 40 is 85%. Therefore, we believe that using obituaries as a data source did not introduce any selection bias with respect to parity.

However, there are several limitations with our study. First, the amount of detail provided in obituaries is fairly low for more in-depth analysis. Obituary content cannot be used to derive information about additional factors such as ethnicity, socio-economic status, age-at-first pregnancy, breastfeeding, and lifestyle choices (e.g., smoking, exercise), which are all known to play a role in cancer risk and therefore should be used to adjust the derived odds ratios accordingly. This is a serious limitation for carrying out a more in-depth analysis from obituary data. Second, the risk of selection bias is always increased with online sources. For example, we observed that cancer is mentioned more openly as the cause of death in obituaries from Utah than New York suggesting geographical differences in terms of people's openness in sharing such details. In addition, cancer was mentioned more often as the cause of death in the obituaries of younger than older people, particularly for breast cancer (as shown in Table 2). The latter is a potential source for information bias. However, this information bias is likely similar among nulliparous and parous women, resulting in non-differential misclassification of the cancer status and bias towards the null. Depending on the question of interest, applying creative correction factors could be useful to mitigate the risk of selection and information bias. Last but not least, the risk of inference error with big but superficial data is typically very high. Online sources may be a

rich resource to collect in an automated, cost-effective way a very large number of candidate subjects, however the quality of the collected information is variable and often very low due to the inherent limitations of natural language processing. Therefore, the process and criteria applied for declaring impact should be far more rigorous since minute differences and wrong findings become easily statistically significant with large sample sizes. This is a well-known challenge for the big health data community (58,59). Based on our findings, comparing and contrasting the trends across the various cancer types was informative and less prone to inferential error in terms of the *relative* impact of parity on cancer risk. For our study, the "law of large numbers" helped in the sense that the quality flaws of individual subject profiles did not compromise the reliability of the relative trends emerging from the data.

In conclusion, the large online content collected and the sophisticated cyber-informatics tools developed to harvest the collected content led us to general findings similar to those produced with traditional epidemiological studies. The study demonstrated that online information sources when leveraged carefully in new and creative ways enable cost-effective and reliable exploratory epidemiological investigations. The main strength of the general cyber-informatics approach we presented lies in its efficiency and ability to monitor trends in a dynamic way by continuously parsing and analyzing emerging open online content. However, given the discussed limitations the proposed approach should be viewed as a promising strategy to generate interesting study hypotheses for guiding and prioritizing future epidemiological studies.

ACKNOWLEDGEMENT

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 study was supported by the National Cancer Institute at the National Institutes of Health (Grant No. 1R01-CA170508).

REFERENCES

1. Lefebvre RC, Bornkessel AS. Digital social networks and health. *Circulation*. 2013;127(17):1829-1836.

2. Chretien KC, Kind T. Social media and clinical care ethical, professional, and social implications. *Circulation*. 2013;127(13):1413-1421.

3. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am J Prev Med*. 2011;40(5)(suppl 2):S154-158.

4. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *J Med Internet Res*. 2013;15(7):e147.

5. Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Arch Med Res*. 2005;36:724-730.

6. Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, et al. Information technology and global surveillance of cases of 2009 H1N1 influenza. *N Engl J Med*. 2010;362:1731-1735.

7. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res*. 2008;10(3):e22.

8. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *J Med Internet Res*. 2010;12(1):e4.

9. Wong PW, Fu KW, Yau RS, Ma HH, Law YW, Chang SS, Yip PS. Accessing suicide-related information on the internet: a retrospective observational study of search behavior. *J Med Internet Res.* 2013;15(1):e3.

10. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. *J Am Med Inform Assoc.* 2014;pii:amiajnl-2014-002841.

11. Brigo F, Lochner P, Tezzon F, Nardone R. Web search behavior for multiple sclerosis: An infodemiological study. *Multiple Sclerosis and Related Disorders*. 2014;3(4): 440-443.

12. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-related hot topic detection in online communities using text clustering. *PloS One.* 2013;8(2):e56221..

13. Cugelman B, Thelwall M, Dawes P. Online interventions for social marketing health behavior change campaigns: a meta-analysis of psychological architectures and adherence factors. *J Med Internet Res*. 2011;13(1):e17.

14. Ayers JW, Althouse BM, Allem JP, Ford DE, Ribisl KM, Cohen JE. A novel evaluation of World No Tobacco day in Latin America. *J Med Internet Res*. 2012;14(3):e77.

15. Ayers JW, Ribisl K, Brownstein JS. Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 "SCHIP" cigarette tax increase. *PLoS One*. 2011;6(3):e16777..

16. Wicks P, Vaughan TE, Massagli MP, Heywood J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*. 2011;29:411-414.

17. Frost J, Okun S, Vaughan T, Heywood J, Wick P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. *J Mel Internet Res*. 2011;13(1):e6.

21. Portier K, Greer GE, Rokach L, Ofek N, et al. Understanding topics and sentiment in an online cancer survivor community. *JNCI Monographs*. 2013;(47):195-198.

22. Kim E, Han JY, Moon TJ, et al. The process and effect of supportive message expression and reception in online breast cancer support groups. *Psychooncology.* 2012;21(5):531-540.

23. DeVita VT Jr, Hellman S, Rosenberg SA. *Cancer: principles and practice of oncology*. New York, NY: Lippincott Williams & Wilkins; 2012.

24. Pathak DR, Speizer FE, Willett WC, Rosner B, Lipnick RJ. Parity and breast cancer risk: possible effect on age at diagnosis. *Int J Cancer*. 1986;37:21-25.

25. Layde PM, Webster LA, Baughman AL, Wingo PA, Rubin GL, Ory HW, and the Cancer and Steroid Hormone Study Group. The Independent associations of parity, age, at first full term pregnancy, and duration of breastfeeding with the risk of breast cancer. *J Clin Epidemiol*. 1989;42(10):963-973.

26. Kelsey JL, Gammon MD, John EM. Reproductive factors and breast cancer. *Epidemiol Rev*. 1993;15:36-47.

27. Beral V, Reeves G. Childbearing, oral contraceptive use, and breast cancer. *Lancet* 1993;341:1102.

28. Lambe M, Hsieh C, Chan H, Ekbom A, Trichopoulos D, Adami HO. Parity, Age at first and last birth, and risk of breast cancer: A population-based study in Sweden. *Breast Cancer Res Treat*. 1996;38:305-311.

29. Möller T, Olsson H, Ranstam J, & Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47

epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. *Lancet*. 2002;360(9328):187-195.

30. Woolcott CG, Koga K, Conroy SM, et al. Mammographyc Density, Parity and age at first birth, and risk of breast cancer: An analysis of four case-control studies. *Breast Cancer Res Treat*. 2012; 132(3):1163-1171.

31. Braem MG, Onland-Moret NC, van den Brandt PA, Goldbohm RA, Peeters PH, Kruitwagen RF, et al. Reproductive and hormonal factors in association with ovarian cancer in the Netherlands cohort study. *Am J Epidemiol*. 2010;172:1181-1189.

32. Moorman PG, Calingaert B, Palmieri RT, Iversen ES, Bentley RC, Halabi S, et al. Hormonal risk factors for ovarian cancer in premenopausal and postmenopausal women. *Am J Epidemiol*. 2008; 167:1059-1069.

33. Tung KH, Goodman MT, Wu AH, McDuffie K, Wilkens LR, Kolonel LN, et al. Reproductive factors and epithelial ovarian cancer risk by histologic type: a multiethnic case-control study. *Am J Epidemiol*. 2003;158:629-638.

34. Tsilidis KK, Allen NE, Key TJ, et al. Oral contraceptive use and reproductive factors and risk of ovarian cancer in the European Prospective Investigation into Cancer and Nutrition. *Br J Cancer*. 2011;105(9):1436-1442.

35. Whittemore A, Harris R, Itnyre J. Collaborative Ovarian Cancer Group. Characteristics relating to ovarian cancer risk: Collaborative analysis of 12 US case-control studies. *Am J Epidemiol*. 1992; 136(4):1184-1203.

36. Le DC, Kubo T, Fujino Y, Sokal DC, Vach TH, Pham TM, Matsuda S. Reproductive factors in relation to ovarian cancer: a case–control study in Northern Vietnam. *Contraception*. 2012;86(5):494-499.

37. Pasalich M, Su D, Binns CW, Lee AH. Reproductive factors for ovarian cancer in southern Chinese women. *J Gynec Oncol*. 2013;24(2):135-140.

38. Titus-Ernstoff L, et al. Menstrual and reproductive factors in relation to ovarian cancer risk. *Br J Cancer*. 2001;84(5): 714-721.

39. Vachon CM, et al. Association of parity and ovarian cancer risk by family history of breast or ovarian cancer in a population-based study of postmenopausal women. *Epidemiology*. 2002;13(1): 66-71.

40. Moorman PG, Palmieri RT, Akushevich L, Berchuck A, Schildkraut JM. Ovarian cancer risk factors in African-American and white women. *Am J Epidemiol.* 2009;170:598–606.

41. Wahi MM, Shah N, Schrock CE, Rosemurgy AN, Goldin SB. Reproductive factors and risk of pancreatic cancer in women: a review of the literature. *Ann Epidemiol*. 2009;19: 103–111.

42. Guan HB, Wu L, Wu QJ, et al. Parity and pancreatic cancer risk: a dose-response meta-analysis of epidemiologic studies. *PLoS One*. 2014;9(3):e92738.

43. Dahabreh IJ, Trikalinos TA, Paulus JK. Parity and risk of lung cancer in women: Systematic review and meta-analysis of epidemiological studies. *Lung Cancer*. 2012;76(2):150-58.

44. Lo AC, Soliman AS, Khaled HM, Aboelyazid A, Greenson JK. Lifestyle, occupational, and reproductive factors and risk of colorectal cancer. *Dis Colon Rectum*. 2010;53:830–837.

45. Peters RK, Pike MC, Chang WWL, Mack TM. Reproductive factors and colon cancers. *Br J Cancer*. 1990;61:741-748.

46. Martinez ME, Grodstein F, Giovannucci E, Colditz GA, Speizer FE, et al. A prospective study of reproductive factors, oral contraceptive use, and risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev*. 1990; 6:1-5.

47. Bostick RM, Potter JD, Kushi LH, Sellers TA, Steinmetz KA, et al. Sugar, meat, and fat intake, and non-dietary risk factors for colon cancer incidence in Iowa women (United States). *Cancer Causes Control*. 1994;5:38-52.

48. Zervoudakis A, Strickler HD, Park Y, Xue X, Hollenbeck A, et al. Reproductive history and risk of colorectal cancer in postmenopausal women. *J Natl Cancer Inst*. 2011;103: 826–834.

49. Wernli KJ, et al. The relationship between gravidity and parity and colorectal cancer risk. *J Women's Health*. 2009;18(7):995-1001

50. Tsilidis KK, Allen NE, Key TJ, Bakken K, Lund E, et al. Oral contraceptives, reproductive history and risk of colorectal cancer in the European Prospective Investigation into Cancer and Nutrition. *Br J Cancer*. 2010;103:1755-1759.

51. Troisi R, Schairer C, Chow WH, Schatzkin A, Brinton LA, et al. Reproductive factors, oral contraceptive use, and risk of colorectal cancer. *Epidemiology*. 1997;8:75-79.

52. Guan HB, Wu QJ, Gong TT, Lin B, Wang YL, Liu CX. Parity and risk of colorectal cancer: a dose-response meta-analysis of prospective studies. *PLoS One*. 2013;8(9):e75279.

53. Xu S, Yoon HJ, Tourassi GD. A user-oriented web crawler for selectively acquiring online content in e-health research. *Bioinformatics*. 2014;30(1):104-114.

54. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstration*. 2014;55-60.

55. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

56. Siegel RL, Miller KD, Jemal !. Cancer Statistics, 2015. *CA Cancer J Clin.* 2015;65:5–29.

57. Martinez G, Daniels K, Chandra A. Fertility of men and women aged 15–44 years in the United States: National Survey of Family Growth, 2006-2010. *Natl Health Stat Report*. 2012; 2010(2006):1-28.

58. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clinical and Translational Science*. 2014;7(4): 342-346.

59. Kaplan RM, Riley WT, Mabry PL. News from the NIH: leveraging big data in the behavioral sciences. *Translational Behavioral Medicine* 2014;4(3):229-231.