Published in final edited form as:

Annu ORNL Biomed Sci Eng Cent Conf. 2013 May ; 2013: 1–4. doi:10.1109/BSEC.2013.6618493.

A cost-effective, case-control study on the association between breast cancer and pregnancy through web mining

Hong-Jun Yoon [Member, IEEE], Songhua Xu [Member, IEEE], and Georgia Tourassi [Member, IEEE]

Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA

Abstract

We report a case-control, breast cancer epidemiological study through mining people stories from the Internet. The aim of the study is to test whether mining openly available, personal stories from the Internet can be a cost-effective way for reliable epidemiological discoveries. As a case study, we focus on the association between breast cancer risk and pregnancy, which is clearly established through controlled clinical survey studies. Specifically, we mined 30,000 online obituary articles. Replicating a case-control study design, our web mining based approach confirmed the general trends reported by traditional epidemiological studies. Our web mining study demonstrates promising preliminary evidence that online content mining can be a cost-effective way for epidemiological knowledge discovery.

I. Introduction

Case-control epidemiological studies are timeconsuming, labor-intensive, and costly. This is particularly true for discovering risk factors of diseases with long-latency such as cancer. To address these challenges inherent to traditional epidemiological studies, we aimed to test an alternative way for conducting epidemiological investigations by leveraging patient data publicly available on the Internet. The purpose of our study is to explore whether human subject records collected through crawling online patient content can enable a new way of Internetbased epidemiological investigation that produces reliable findings at a significantly reduced cost and time.

As a case study we focus on the association between breast cancer risk and pregnancy. Prior studies have shown that a woman's exposure to ovarian hormones affect her risk of developing breast cancer throughout her lifetime [1]. Pregnancy and breastfeeding reduce the number of a woman's menstrual cycles, which results in reduced cumulative hormonal exposure and thus lower breast cancer risk. Large-scale, case-control, observational studies have clearly established the relationship between the breast cancer risk and pregnancy [2-5]. We aim to discover the same relationship via web mining.

II. Methodology

We mined obituary articles from online newspapers since such material is widely available and typically includes content regarding a person's family members and cause of death. The language of obituary articles is also often standardized, facilitating automatic computer parsing.

Via web crawling we collected obituary articles from local newspapers published online between 2000 and 2012 in the United States. From each obituary, we automatically extracted the deceased subject's longevity, gender, and offspring information. Those obituary articles from which we could not identify the subject's age were excluded in our study. Gender identification was relatively easy to accomplish. Since male breast cancer is not of interest in this study, obituaries of males were also excluded in our analysis. History of childbirth was inferred by the listing of surviving offsprings mentioned in an obituary. For those offsprings clearly stated as stepchildren or adopted, we did not count them as biological children. History of breast cancer was inferred from explicit statements that breast cancer was the cause of death.

After parsing obituary articles through the above procedure, we formed a case group with breast cancer subjects. We also compiled another "control" group that consists of all female subjects for whom longevity information was extractable by our text parser. Obituaries of females for whom breast cancer was not explicitly stated were included in the control group. Therefore, the control group may have included breast cancer patients or survivors because. Although this was a limitation, it could only influence negatively our findings by masking any difference between breast cancer patients and non-breast cancer female subjects.

III. Results

The total numbers of obituary articles for the control group and for breast cancer group were respectively 15,954 and 20,332. The number of subjects with at least one biological child among the control group and the case (disease) group were 13,548 and 15,946 respectively. That is, 84.9% of control subjects gave childbirth while only 78.4% of case subjects did. The difference is statistically significant according to the difference of two binomial proportions statistical test. The 95% confidence interval of the difference is [5.7%, 7.3%]. Figure 1 shows the relative population size for each age group.

We further compared the childbirth incidence rate in the two groups according to the subjects' age at death. Figure 2 reports the respective percentages. To derive the confidence intervals for each age group, we applied bootstrapping. Figure 2 shows that for the older groups (i.e., subjects who died after the age of 70), there was no significant difference in childbirth incidence between the control and the disease groups. This finding is expected since age is a stronger breast cancer risk factor. Similarly, for the youngest group (i.e., those younger than 30 yrs old), there was no significant difference in childbirth incidence. The lack of any meaningful finding is expected due to the early death of the subjects. Therefore, what is most interesting is the comparison between the two groups between the ages of 30 to 69yrs old, for which significant differences were observed between the control and the disease group in terms of the percentages of subjects who gave childbirth. Figure 2 confirms

the traditional epidemiological understanding that women in the disease group are much less likely to have given childbirth than women in the control group.

IV. Discussion

This paper reports a cost-effective, case-control study on breast cancer epidemiology by mining a large number of patient records from the Internet. We do recognize that our current approach does not derive further details regarding a subject's key life events such as the age of her first pregnancy or breastfeeding experience, which limits us from carrying out more in-depth analysis. Such information may never be expected from obituaries. In our future study, we plan to explore other types of online patient data regarding a person's life stories for more comprehensively and robustly conducting web mining based epidemiological explorations.

Acknowledgments

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.

References

1. Epstein FH. Estrogen and the Risk of Breast Cancer. *NEJM*. Jan 25; 2001 344(4):276–286. [PubMed: 11172156]
2. Ramon JM, Escriba JE, Casas I, Benet J, Iglesias C, Gavalda L, Torras G, Oromi J. Age at First Full-Term Pregnancy, Lactation and Parity and Risk of Breast Cancer: A Case-Control Study in Spain. *Eur J Epidemiol*. 1996; 12:449–453.
3. Lambe M, Hsieh C, Chan H, Ekblom A, Trichopoulos D, Adami H. Parity, Age at First and Last Birth, and Risk of Breast Cancer: A Population-based Study in Sweden. *Breast Cancer Research and Treatment*. 1996; 38:305–311. [PubMed: 8739084]
4. Enger SM, Ross RK, Handerson B, Bernstein L. Breastfeeding History, Pregnancy Experience and Risk of Breast Cancer. *Br J Cancer*. 1997; 76(1):118–123. [PubMed: 9218743]
5. Lord SJ, Bernstein L, Johnson KA, et al. Breast Cancer Risk and Hormone Receptor Status in Older Women by Parity, Age of First Birth, and Breastfeeding – A Case-Control Study. *Cancer Epidemiol Biomarkers Prev*. Jul; 2008 17(7):1723–1730. [PubMed: 18628424]

Figure 1.
Relative population size for each age group in the control and diseased groups.

Figure 2.
Childbirth incidence rate in the two comparison groups.