Knowledge Engineering Framework to Quantify Dependencies between Epidemiological and Biomolecular Factors in Breast Cancer

Iuliia Innokenteva[1], Richard Hammer[2,1] and Dmitriy Shin[2,1]

[1] *MU Informatics Institute,* [2]*Department of Pathology and Anatomical Sciences*
University of Missouri, 1 Hospital Dr. Pathology, Med Sci Bldg,
Columbia, MO, 65203, USA
**Email: shindm@health.missouri.edu*

Abstract. The relationship between social determinants of health (SDoH) and chronic disease risks is crucial for its prevention. Such associations are relatively easier to uncover for simple diseases such as obesity or heart diseases. But for complex diagnoses like cancer, a large number of factors contribute to the onset of the disease. For instance, there is increasing evidence that biomolecular factors of cancer can be influenced by behavioral and environmental patterns. For example, several subtypes of breast cancer that respond to different hormonal therapies can arise due to different lifestyle, social, physiological risk-factors. Cancer Registries and EHRs as the sources of health data are used widely in epidemiological research. Being collected by health professional, the EHR data reduce research cost and embraces the whole population. However, the primary purpose of those records is not being used in a research. Therefore, data adjusting issue can arise. Often the structure of records is not satisfying to build an epidemiological model. To fit data from EHR and Cancer Registry to epidemiological model we propose the method of knowledge engineering to construct Bayesian Networks (BN) structure using control vocabularies. Specifically, we selected fields from records and used National Institute of Cancer Thesaurus to determine nodes for BN structure. We demonstrate utility of this approach on a cohort of University of Missouri Hospital (UMH) patients who was diagnosed with breast cancer.

Keywords: Breast Cancer, Epidemiology, Controlled Vocabulary, Ontology, Bayesian Network, Knowledge Engineering, Biomolecular factors, Hormone Receptors.

1 Introduction

Recently, EHRs has been used largely as a source of health data for epidemiological research. Readily available data collected in accordance with health facilities' standards help reducing research costs and saving time. Systematic review made by Casey et al. shows that extract, transform, load (ETL) tool is mainly used to make health data suitable for researchers (Casey, et al., 2016). Different common data models (CDM) such as Observational Medical Outcomes Partnership (OMOP), FDA Sentinel Initiative, and the Patient Centered Outcome Research Network (PCORNet) are based on ETL approach (Califf, 2014; Carnahan, et al., 2014; Kahn, et al., 2012). CDMs listed above

aim to integrate and adjust health data from diverse sources such as health care providers, pharmacies, laboratories, etc. The adjustment part of CDM technique is to bring the data to the same consistent format using controlled vocabularies (e.g., same variable names, attributes, etc.) (Resnic, et al., 2015). However, it is not well suited for the selection of pertinent variables to design specific research studies, especially in the domain of complex diseases, such as cancer.

According to the World Health Organization (WHO), one-third of all cancer cases can be prevented by having dietary changes, stopping from smoking, getting hepatitis vaccinations, and exercising regularly. Breast cancer is the most commonly diagnosed cancer worldwide and particularly in the USA (WHO, 2016). Still, it is one of the cancer types which can be partially prevented by lifestyle modification (CDC, 2018). There are specific subtypes of breast cancer, which are characterized by different hormonal patterns. The most commonly used at breast cancer diagnostic and treatment hormone receptors are estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). In accordance with their presence or absence in a body, breast cancer is divided into subtypes. For instance, luminal cancer tends to be ER negative, basal-like breast cancer is usually triple negative and it is the most challenging type of the disease.

Association between lifestyle factors and those two breast cancer subtypes is shown in Butler et al study. According to it, smoking is positively associated with luminal cancer and almost does not effect basal-like cancer (Butler et al., 2016). Similar study considered obesity as a risk-factor and a significant association between triple negative breast cancer and obesity was found (Turkoz, et al., 2013). Smoking and ER-positive cancer analysis showed that current smokers are more susceptible for ER-positive breast cancer (Odds Ratio [OR]=1.6) than ever smokers (OR=1.4). But there was no difference in terms of triple negative cancer risk in both groups (Kawai, Malone, Tang, & Li, 2014). Statistical evidence of associations between obesity and ER-positive cancer was proven in Nechuta et al. study. The same research showed strong correlation between alcohol consumption and ER-positive breast cancer (Nechuta, et al., 2016). However, some studies presented that higher body mass index increases breast cancer risk independently on menopausal status and estrogen receptor (ER) expression (Schirer, et al., 2013; Wada et al., 2014). Another interesting finding is that urban women have higher incidence rates (IRR) of ER-positive breast cancer (IRR=3.36) than rural women (Dey, et al., 2009). After reviewing literature described above, we have determined potential risk factors for all subtypes of breast cancer. Smoking, alcohol consumption, obesity had been chosen as initial variables for our research. Additionally, we considered the most common comorbidities such as hypertension and diabetes as risk-factors.

Combination of molecular biology approaches and epidemiology studies can help to determine the causes of certain subtype of breast cancer. Bayesian Networks can be instrumental to model such processes. BN is a graphical model that represents relationships between factors and their probabilities. The model is usually used for prediction of disease risk depending on certain factors (Rosa, et al., 2015). Each variable is represented as a node of BN and it has several mutually exclusive instances. Changing instances for independent variables and setting a dependent variable as a target we can predict an outcome.

Still, it is not a trivial process to select appropriate entities from a EHR to determine nodes for a BN model. Specifically, there has to be a protocol to determine appropriate level of granularity for those entities. For instance, several fields in EHR system might have to be aggregated to represent a node in BN model.

To address this problem, we aim to create a knowledge engineering framework utilizing controlled vocabularies such as ontologies and thesauri. Determined through such a process BN nodes are then connected in a structure to compute conditional probabilities. Then the BN model can be used to quantify and predict factors that influences hormonal patterns of breast cancer, which can lead to better patient care.

2 Methods

The pipeline of knowledge engineering process is shown on Figure 1.

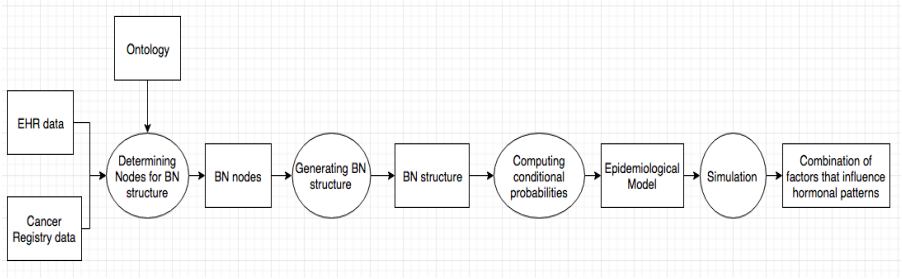

Fig. 1. Knowledge engineering process for prediction of breast cancer hormonal patterns.

Data are selected from EHR and Cancer registry based on epidemiological knowledge about breast cancer. There is number of possible factors contributing to an onset of breast cancer including demographic, socio-economic, physiological, and mental factors. For the given research we used risk factors that were available in UMH EHRs and Cancer Registry records. Ontology is used to determine which EHR fields can be aggregated. We used the National Cancer Institute (NCI) Thesaurus to select potential cancer risk factors that later could be retrieved from EHR. For example, according to NIC Thesaurus, variables 'Type 1 Diabetes Mellitus' and 'Type 2 Diabetes Mellitus' are the child concepts of 'Diabetes Mellitus' concept. Thus, depending on epidemiological context those variables can be aggregated in one BN node "Diabetes Mellitus' with possible values 'Type 1', 'Type 2', 'Undefined Diabetes', 'No History of Diabetes'.

To generate the BN structure, we used epidemiological knowledge and literature review presented in the introduction. In addition to history of obesity, tobacco and alcohol consumption we included comorbidities such as diabetes and hypertension. We added a race variable as well to make the causality pattern more representative. Generated BN structure and its nodes are shown in Figure 2.

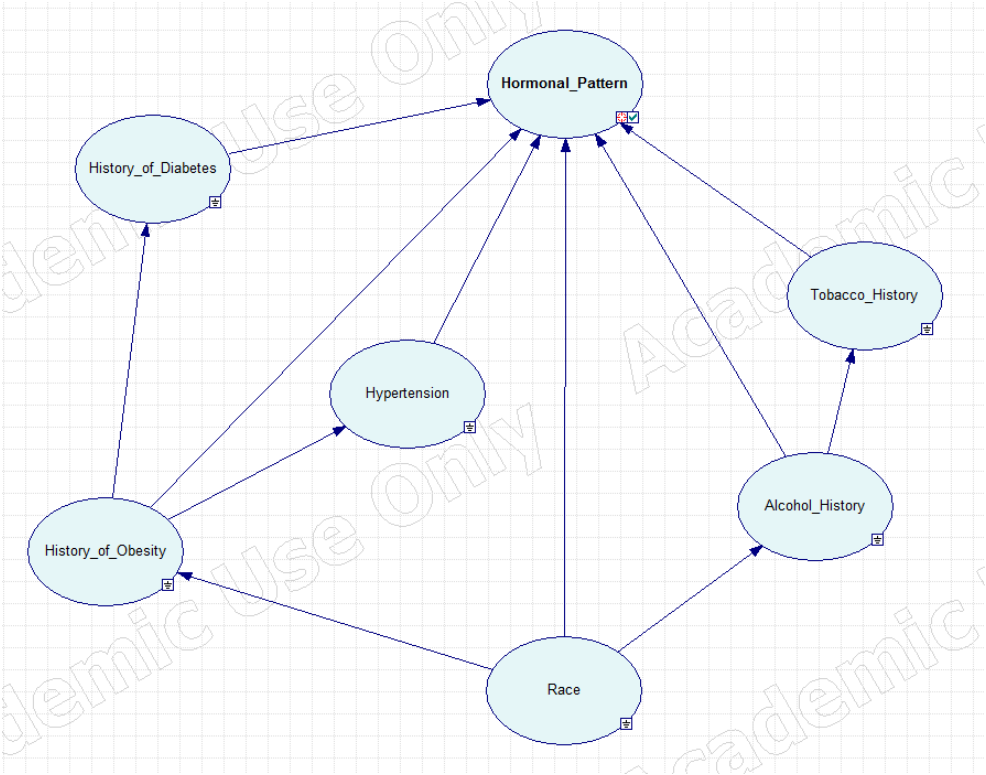

Fig. 2. Expert-based BN structure

For the generated structure, we learned parameters from the dataset of 980 patients of UMH with diagnosed breast cancer. Then setting 'Hormonal_Pattern' node as a target and setting different values for other nodes we could simulate cases and predict hormonal pattern of breast cancer depending on behavioral, health, and social factors (Figure 3).

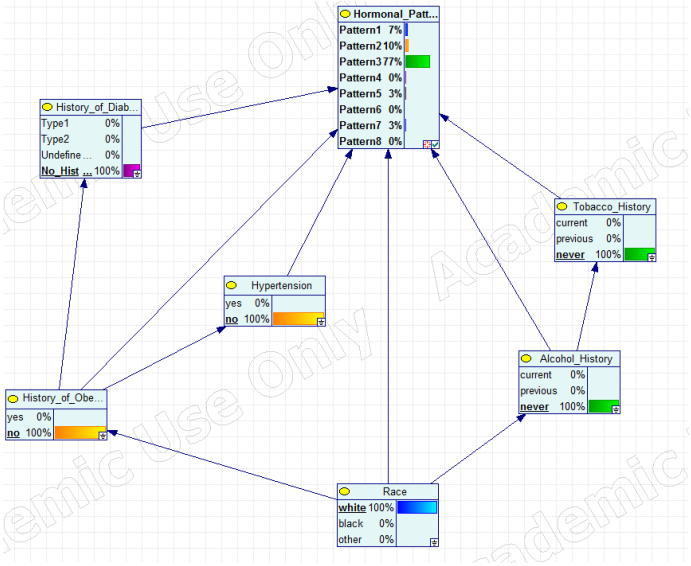

Fig. 3. Example of simulation

3 Results and Discussion

Using UMH Cancer Registry data we determined a cohort of 1070 patients who were diagnosed with breast cancer after 2013. Information about race, history of tobacco and alcohol use, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) was found from cancer registry data. Information about history of obesity, diabetes, and hypertension was added from UMH EHRs. Hormonal patterns were defined as eight combinations of ER, PR, HER2 different values, positive or negative (Table 1). During the data cleaning process 90 cases were removed because of missing values.

Table 1. Combinations of hormonal patterns

Hormonal Pattern	ER value (+/-)	PR value (+/-)	HER2 value (+/-)
1	+	+	+
2	+	-	-
3	+	+	-
4	+	-	+
5	-	-	-
6	-	+	+
7	-	-	+
8	-	+	-

Table 2 contains randomly selected five cases with different values of nodes. The results of simulation for five given cases are presented on the Table 2 as probabilities of different combinations of ER, PR, HER2 values.

Table 2. Cases for simulation process

Case#	Race	History_of_ Obesity	History_of_ Diabetes	Hyperten-sion	Alcohol History	Tobacco History
1	Black	Yes	Type 2	Yes	Current	Current
2	White	No	No_History	No	Never	Never
3	Black	No	No_History	No	Never	Never
4	Black	No	Type 1	Yes	Never	Previous
5	White	Yes	Type 2	Yes	Current	Current

Table 3. Probabilities of hormonal pattern for simulation cases

Case#	Probability of Pattern1,%	Probability of Pattern2,%	Probability of Pattern3,%	Probability of Pattern4,%	Probability of Pattern5,%	Probability of Pattern6,%	Probability of Pattern7,%	Probability of Pattern8,%
1	50	50	0	0	0	0	0	0
2	7	10	77	0	3	0	3	0
3	0	31	56	0	8	0	5	0
4	13	11	0	0	76	0	0	0
5	50	50	0	0	0	0	0	0

The results of simulating different cases with certain values show that some variables influence more than others on the 'hormonal pattern' outcome. Changing values one by one, we can see which of the nodes has the major effect on hormone receptors pattern. This approach can be used to predict a risk of certain subtype of breast cancer depending on a variety of factors. In a best-case scenario, we could predict triple negative breast cancer risk which is the most challenging subtype of the disease in terms of response for a therapy.

Using the knowledge engineering pipeline presented in the study, one can add variables from different sources and aggregate them using ontology. For instance, EHRs contain patients' addresses and it can be useful source of information in epidemiological sense. The thesaurus has a class called 'Group' which is then divided into 'rural/underserved population' and 'urban population'. To extract this useful information, the nominal 'address' variable from the EHR needs to be modified to rural/urban categorical variable. Then it can be included in epidemiological model to predict breast cancer subtype depending on patients' residency which represents an access to health care.

For the epidemiological model of breast cancer hormonal patterns, we did not include all possible predictors of the disease such as age, marital status, age at menarche, age

at menopause, number of pregnancies. The purpose of the model is to show the possibility of utilizing the pipeline for certain population health research.

Future research can be done to validate the results of this study. Using data analysis statistical tools such as STATA one can analyze associations between nodes and find evidence of statistical significance.

4 Conclusion

To address the problem of determining the granularity of the data entities from different sources, we created a knowledge engineering pipeline. By utilizing the pipeline, we could modify types of information from EHR and Cancer Registry records using a controlled vocabulary such as NIC Thesaurus. We converted those variables into useful for epidemiological models form. The utilization of this pipeline is not limited by cancer epidemiology purposes only. It can be used for other population health research aimed to study health care access, behavioral patterns, treatment or public health program effectiveness, and many other aspects.

References

Butler, E. N., Tse, C.-K., Bell, M. E., Conway, K., Olshan, A. F., & Troester, M. A. (2016). Active smoking and risk of Luminal and Basal-like breast cancer subtypes in the Carolina Breast Cancer Study. *Cancer Causes & Control : CCC*. https://doi.org/10.1007/s10552-016-0754-1

Califf, R. M. (2014). The Patient-Centered Outcomes Research Network. *North Carolina Medical Journal,75*(3), 204-210. doi:10.18043/ncm.75.3.204

Carnahan, R. M., Bell, C. J., & Platt, R. (2014). Active Surveillance: The United States Food and Drug Administrations Sentinel Initiative. *Manns Pharmacovigilance,*429-437. doi:10.1002/9781118820186.ch2

Casey, J. A., Schwartz, B. S., Stewart, W. F., & Adler, N. E. (2016). Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annual Review of Public Health,36*, 61-81.

Centers for Disease Control and Prevention. Breast Cancer. (2018, May 22). Retrieved from https://www.cdc.gov/cancer/breast/index.htm

Dey, S., Soliman, A. S., Hablas, A., Seifeldin, I. A., Ismail, K., Ramadan, M., . . . Merajver, D. (2009, June 23). Urban–rural differences in breast cancer incidence by hormone receptor status across 6 years in Egypt. Retrieved from https://link.springer.com/article/10.1007/s10549-009-0427-9, https://doi.org/10.1007/s10549-009-0427-9

Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data Model Considerations for Clinical Effectiveness Researchers. *Medical Care,50*. doi:10.1097/mlr.0b013e318259bff4

Kawai, M., Malone, K. E., Tang, M.-T. C., & Li, C. I. (2014). Active smoking and the risk of estrogen receptor-positive and triple-negative breast cancer among women ages 20 to 44 years. *Cancer*, *120*(7), 1026–1034. https://doi.org/10.1002/cncr.28402

Nechuta, S., Chen, W. Y., Cai, H., Poole, E. M., Kwan, M. L., Flatt, S. W., … Ou Shu, X. (2016). A pooled analysis of post-diagnosis lifestyle factors in association with late estrogen-receptor-positive breast cancer prognosis. *International Journal of Cancer*. https://doi.org/10.1002/ijc.29940

Resnic, F., Robbins, S., Denton, J., Nookala, L., Meeker, D., Ohno-Machado, L., . . . Fitzhenry, F. (2015). Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. *Applied Clinical Informatics,06*(03), 536-547. doi:10.4338/aci-2014-12-cr-0121

Rosa, C. M. I., Simões, P. W., Doneda, G., Silva, D., Moretti, G. P., Simon, C. S., … Rosa, M. I. (2015). Meta analysis of the use of Bayesian networks in breast cancer diagnosis. *Cad. Saúde Pública*, *31*(311), 26–3826. https://doi.org/10.1590/0102-311X00205213

Schairer, C., Li, Y., Frawley, P., Graubard, B. I., Wellman, R. D., Buist, D. S. M., … Miglioretti, D. L. (2013). Risk factors for inflammatory breast cancer and other invasive breast cancers. *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/djt206

Turkoz, F. P., Solak, M., Petekkaya, I., Keskin, O., Kertmen, N., Sarici, F., … Altundag, K. (2013). Association between common risk factors and molecular subtypes in breast cancer patients. *The Breast*, *22*(3), 344–350. https://doi.org/10.1016/J.BREAST.2012.08.005

Wada, K., Nagata, C., Tamakoshi, A., Matsuo, K., Oze, I., Wakai, K., … Research Group for the Development and Evaluation of Cancer Prevention Strategies in Japan. (2014). Body mass index and breast cancer risk in Japan: a pooled analysis of eight population-based cohort studies. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO*. https://doi.org/10.1093/annonc/mdt542

World Health Organization. Breast cancer: Prevention and control. (2016, January 21). Retrieved from http://www.who.int/cancer/detection/breastcancer/en/