An Exploratory Analysis of a Large Health Cohort Study Using Bayesian Networks

By Delin Shen

S.B. Biomedical Engineering, S.B. Mechanical Engineering
Tsinghua University, 1994
S.M. Biomedical Engineering
Tsinghua University, 1997

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQURIEMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2006

Signature of Author……...……...
Harvard-MIT Division of Health Sciences and Technology
January 26, 2006

Certified by……...……
Peter Szolovits, Ph.D.
Professor of Computer Science and Engineering
Professor of Health Sciences and Technology
Thesis Supervisor

Accepted by……...……...
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Co-Director, Harvard-MIT Division of Health Sciences and Technology

An Exploratory Analysis of a Large Health Cohort Study Using Bayesian Networks

By Delin Shen

Submitted to the Harvard-MIT Division of Health Sciences and Technology on January 26, 2006
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Health Sciences and Technology

ABSTRACT

Large health cohort studies are among the most effective ways in studying the causes, treatments and outcomes of diseases by systematically collecting a wide range of data over long periods. The wealth of data in such studies may yield important results in addition to the already numerous findings, especially when subjected to newer analytical methods.

Bayesian Networks (BN) provide a relatively new method of representing uncertain relationships among variables, using the tools of probability and graph theory, and have been widely used in analyzing dependencies and the interplay between variables. We used BN to perform an exploratory analysis on a rich collection of data from one large health cohort study, the Nurses' Health Study (NHS), with the focus on breast cancer.

We explored the data from the NHS using BN to look for breast cancer risk factors, including a group of Single Nucleotide Polymorphisms (SNP). We found no association between the SNPs and breast cancer, but found a dependency between clomid and breast cancer. We evaluated clomid as a potential risk factor after matching on age and number of children. Our results showed for clomid an increased risk of estrogen receptor positive breast cancer (odds ratio 1.52, 95% CI 1.11-2.09) and a decreased risk of estrogen receptor negative breast cancer (odds ratio 0.46, 95% CI 0.22-0.97).

We developed breast cancer risk models using BN. We trained models on 75% of the data, and evaluated them on the remaining. Because of the clinical importance of predicting risks for Estrogen Receptor positive and Progesterone Receptor positive breast cancer, we focused on this specific type of breast cancer to predict two-year, four-year, and six-year risks. The concordance statistics of the prediction results on test sets are 0.70 (95% CI: 0.67-0.74), 0.68 (95% CI: 0.64-0.72), and 0.66 (95% CI: 0.62-0.69) for two, four, and six year models, respectively.

We also evaluated the calibration performance of the models, and applied a filter to the output to improve the linear relationship between predicted and observed risks using Agglomerative Information Bottleneck clustering without sacrificing much discrimination performance.

Thesis Supervisor: Peter Szolovits, Ph.D.
Title: Professor of Computer Science and Engineering
Professor of Health Sciences and Technology

To my grandmother,

my parents,

and my wife

Acknowledgments

There are many people who have helped me during my thesis research. First and foremost, I would like to thank my thesis supervisor, Professor Peter Szolovits, of the MIT Computer Science and Artificial Intelligence Lab. I am grateful for his continuous guidance in learning and doing research, his careful reading of my thesis, and especially his moral support during my study at MIT. From time to time, I had challenges in my research work, and talking with Peter always helped me to overcome the difficulties, while his sense of humor made it easier for me to deal with the pressure. I would also very much like to thank my mentor and friend, Professor Marco Ramoni, of the Harvard Medical School and Children's Hospital, especially for his tremendous help with my research and the encouragement I received from his own experiences. Marco spent a large amount of time discussing with me the research, going through with me the presentations, without which I could never have finished the thesis work. I am also grateful to Professor Graham Colditz, of the Harvard School of Public Health, whose insights and expertise in breast cancer and the Nurses' Health Study were instrumental in both my research and thesis defense. Finally, I cannot continue without expressing my lasting gratitude to Professor Tommi Jaakkola, of the MIT Computer Science and Artificial Intelligence Lab, and Professor Lucila Machado, of the Harvard Medical School and Brigham Women's Hospital, who gave me invaluable feedbacks and suggestions in how to apply machine learning techniques.

I am grateful to the members of the Channing Lab, who gave me so many help and support during the research. Dr. Karen Corsano helped me to get familiar with the Nurses' Health Study, and always answered my questions promptly. Lisa Li spent tremendous amount of time extracting the data for our research, which is an extra to her already heavy loaded daily work. Dr. David Hunter and Rong Chen provided and helped me understand the genomics data. Marion McPhee helped me on the log-incidence model and provided related data. Their contributions are all crucial to this thesis work. I would also like to thank all MEDG members, past and present, at the MIT Computer Science and Artificial Intelligence Lab, who have each helped in their individual ways. I would especially like to mention Fern DeOliveira, our awesome assistant, for her always being there and prompt help.

I am indebted to my friends and family who have continued to believe in me, and who have kept me sane, entertained, and happy. I would particularly like to thank Alfred Falco, Wei Guo, Jinlei Liu, Jingsong Wu, Chao Wang, Yi Zhang, Hao Wang, Selina Lam, Jianyi Cui, Jing Wang, Minqu Deng, Xingchen Wang, and Song Gao.

This thesis is based upon work supported in part by a HST-MEMP fellowship, the Defense Advanced Research Projects Agency grant F30602-99-1-0509, and the National Institute of Health grant U54LM008748-01.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

Health care is one of the major concerns of modern society, and much effort has been invested in studying the causes, treatments and outcomes of disease. Large health cohort studies are among the most effective, because they follow a large and relatively stable population over a long period of time and systematically collect comparable longitudinal data, sometimes over decades. There are numerous large health cohort studies being carried on, such as the Framingham Heart Study from 1948, The British Doctors Study from 1954, the Dunedin Multidisciplinary Health and Development Study started in 1972, and the Nurses' Health Study since 1976. Such studies have led to numerous important insights and many publications, and they often form the basis for health care recommendations and policies. We suspect that the wealth of data in such studies may yet yield many additional important results, especially when subjected to newer analytical methods.

In order to work efficiently, we focused our analysis on breast cancer, one of the most frequently diagnosed cancers in the US, and one of the motivating diseases for the Nurses' Health Study. We explored the data from the Nurses' Health Study using Bayesian Networks to look for potential risk factors for breast cancer, and also developed and evaluated risk predicting models of breast cancer.

Bayesian Networks provide a relatively new method of representing uncertain relationships among variables, using the tools of probability and graph theory. Such a network allows a concise representation of probabilistic dependencies and is often used to model potential causal pathways. We have used heuristic techniques that automatically induce a Bayesian Network that fits observational data to suggest the likely dependencies in the data. Bayesian Networks have been widely used in analyzing such dependencies and the interplay between variables. In this work, we

have used Bayesian Networks to perform an exploratory analysis on a rich collection of data from one large health cohort study, the Nurses' Health Study [1].

This thesis is organized into several major chapters. We first introduce the background of this research in Chapter 2, including the Nurses' Health Study, a large health cohort study from which we obtained the data, the tools we used to explore the data, a brief literature review of breast cancer models and their evaluation, and a brief introduction to machine learning and Bayesian Networks.

In Chapter 3, we describe the exploratory analysis performed on the first data sets we obtained, including a proof of concept Bayesian Network, dependency analysis among clinical and life-style variables and a small set of genotypic data, and a finding of association between clomid use and breast cancer with specific estrogen receptor status.

Chapter 4 presents a group of risk predicting models for breast cancer based on the same data used to derive a log-incidence model previously published by Colditz *et al*. We built the new models using Bayesian Networks, an approach different from the original log-incidence model. The models were evaluated on discrimination and calibration abilities. We also used agglomerative information bottleneck clustering to filter the prediction results, and achieved improved linear relationship between the predicted risks and observed risks.

Chapter 5 discusses issues we encountered in this work, summarizes lessons learned from the research, and gives some possible future research direction from this work.

Chapter 2 Background

This chapter introduces the background of our research, including a brief introduction to breast cancer and beast cancer models, Bayesian Networks and other machine learning techniques we employed in our research, and the Nurses' Health Study data on which we based our study.

2.1 Nurses' Health Study

The Nurses' Health Study (NHS), established in 1976 by Dr. Frank Speizer and funded by the National Institute of Health, is among the "largest prospective investigations into the risk factors for major chronic diseases in women" [1]. Registered nurses were selected to be followed prospectively because they were anticipated to be able to "respond with a high degree of accuracy to brief, technically-worded questionnaires and would be motivated to participate in a long term study" due to their nursing education [1]. A short questionnaire with health-related questions is sent to the members every two years starting from 1976, and a long questionnaire with food frequency questions is sent every four years starting from 1980. Questions about quality of life were added to questionnaires since 1992. 33,000 blood samples were collected in 1989 and are stored and used in case/control analyses. 2448 blood samples have been genotyped at 47 Single Nucleotide Polymorphism (SNPs) sites [1][2].

With a follow-up rate exceeding 90% [3], the Nurses' Health Study has a very good longitudinal record of phenotypes and clinical and life-style factors, including physical data, health status, life styles, nutritional intake, family history, etc. Questionnaires were sent to the nurses every other year, and the data available to this study are generally from 1976 to 2000, except for those who passed away or left the study for other reasons.

The collected variables can be divided into different categories, and below is an incomplete but indicative list (with the focus on breast cancer).

- general information: year of birth, month of birth, race, height, weight, geographic location, education, parity (number of children), age at first birth, breast feeding history, weight of heaviest child, father's occupation, marital status, husband's education, living status, years worked in operating room, total number of years on night shifts, hours of sleep, sleeping position, snoring, birth weight, breast fed in infancy, and handedness
- clinically related information: smoking status, age at start of smoking, passive smoking, body mass index, waist and hip measurement, age at menopause, type of menopause, post menopausal hormone use, oral contraceptive use, breast cancer history of mother and sisters, total activity score, menstrual cycle age, regularity of period, tan or sunburn, tan or sunburn as a child, moles, lipstick use, breast implant, multi-vitamin use, social and psychological characteristics
- medical history: alcohol dependency problem, aspirin use, tagamet use, tubal ligation, clomid use, tamoxifen use, talcum powder use, hip or arm fracture, TB test, elevated cholesterol, heart disease, high blood pressure, diabetes, cancer report, lung cancer, ovarian cancer, DES treatment, breast cancer diagnosis and type, estrogen receptor and progesterone receptor test
- diet information: different kinds of food intake

Many valuable medical findings have originated from this study, though contemporary machine learning techniques (except logistic regression) were rarely employed. Therefore, we hope that by exploring the data from NHS using modern machine learning tools we can find interesting new medical evidence, e.g. develop a new breast cancer model, or learn helpful experiences from such an exploratory analysis.

2.2 Breast Cancer Models

Breast cancer has been the most frequently diagnosed and the second most deadly cancer in women in the US for many years. In 1998, breast cancer constituted about 30% of all cancers and

caused about 16% of cancer deaths in women. [1] Recently available statistics from the American Cancer Society estimate that more than two hundred thousand newly diagnosed cases and more than forty thousand deaths resulted from breast cancer in 2004. [6]

The effort to understand the risk factors of breast cancer, and one step further, to predict breast cancer risks, can be traced back to the late 60's. [7-11] In early studies, it was first recognized that age at menarche (the onset of menstruation), age at first birth, and age at menopause are three major risk factors for breast cancer. Generally, early age at menarche and late age at menopause are considered to be associated with higher risk of breast cancer, while early first full-term pregnancy is associated with lower risk. Postmenopausal weight, family history, duration of having a menstrual period, age, pregnancy history, and other risk factors were also investigated in later studies. [13-21]

Starting from the early 80's, scientists started to build mathematical models that can predict a probability, or risk, of getting breast cancer. Interestingly enough, most of these research efforts fall into two groups. Some of them focused on the inheritance of breast cancer, building models based on family history and/or genotype data. The others tried to build the model by combining individual risk factors, mostly reproductive variables such as age at menarche, age at menopause, parity, etc.

Ottman *et al*. published a simple model in 1983 that calculates a probability of breast cancer diagnosis for mothers and sisters of breast cancer patients. [24] They used life-table analysis to estimate the cumulative risks to various ages based upon two groups of patients from the Los Angeles County Cancer Surveillance Program, then derived a probability within each decade between ages 20 and 70 for mothers and sisters of the patients, according to the age of diagnosis of the patient and whether the disease was bilateral or unilateral.

Claus *et al*. developed a genetic model to estimate age-specific breast cancer risks for women with at least one relative with breast cancer. [25-26] The data they used were from the Cancer and Steroid Hormone (CASH) Study, a population-based, case-control study evaluating the impact of oral contraceptives on the risk of breast cancer. The model derived risk estimates based on the relative's age at diagnosis and the degree of relationship of the relative(s). Their results showed that women with first-degree relatives who were diagnosed with breast cancer at early ages have very high lifetime risks of breast cancer. They also suggested BRCA1 susceptibility for breast cancer.

Inspired by the discovery of breast cancer susceptibility genes BRCA1 and BRCA2 between 1994 and 1995, risk models were also developed to predict the probability that an individual might be a carrier of a mutant gene, either the known BRCA1 and BRCA2, or a hypothetical unknown gene BRCAu.

Couch *et al*. examined families with at least two breast cancer cases for germline mutations in BRCA1, and built a model with logistic regression, using average age at breast cancer diagnosis, ovarian cancer history, and Ashkenazi Jewish ancestry as risk factors. [29] Shattuck-Eidens *et al*. developed a similar model on a different group of families, but without the limitation to a family history of breast cancer. [30] Using logistic regression, Frank *et al*. identified ovarian cancer, bilateral breast cancer, and age of diagnosis for breast cancer before 40 as predictors for both BRCA1 and BRCA2. [31] Parmigiani *et al*. developed a Bayesian model to evaluate the probabilities that a woman is a carrier of a mutation of BRCA1 and BRCA2 using breast and ovarian cancer history of first and second degree relatives as predictors. [33]

In the literature, there are many more research projects trying to identify carriers of a mutant gene based on family history, than those trying to predict the risk of breast cancer using genotype data. This is probably due to the high cost of genotyping and the unavailability of a reliable genotype

data set in past times. As indicated by Parmigiani, it cost $2400 to test for both BRCA1 and BRCA2 in 1997.[33] Recently, however, scientists started to build models to predict breast cancer risk using BRCA1 and BRCA2. For example, Tyrer *et al*. published a model that incorporated BRCA1, BRCA2, and a hypothetical low penetrance gene, as well as some personal risk factors. [39] No doubt in the foreseeable future we will see more and more models using genotypic data.

In the other group using individual risk factors, Moolgavkar *et al*. proposed one of the earliest risk prediction models in 1980, which predicts age-specific incidence of breast cancer in females, based upon physiologic response of breast tissue to menarche, menopause, and pregnancy on the cellular level. [18] They suggested a two-stage model that incorporates growth of breast tissues to derive an age-specific incidence curve, which can explain with close quantitative agreement the observed risk due to age at menarche, age at menopause, and parity in a combined data set including data from Connecticut, Denmark, Osaka (Japan) Iceland, Finland, and Slovenia. This "tissue aging theory" has been modified and extended in many later research projects.

Pike *et al.* proposed a quantitative description of "breast tissue age" based on age at menarche, first full-term pregnancy, and age at menopause, which fits well a linear log-log relationship between breast cancer incidence and age. [23] The Pike model assumes breast tissue aging started from menarche at a constant f_0, dropped after the first full-term pregnancy to another constant f_1, and then decreased linearly from age 40, which they called the perimenopausal period, to the last menstrual period, and finally kept constant at that level.

In 1989, Gail et al. proposed what is now called the Gail model, a breast cancer risk model clinically applied today, based on the Breast Cancer Detection Demonstration Project (BCDDP). [27] The relative risk of the original model was derived from a matched case-control subset from BCDDP, using unconditional logistic regression on five risk factors: age, age at menarche, age at first live birth, number of previous biopsies, and number of first-degree relatives with breast

cancer. Women in BCDDP regularly received mammographic screening, so special caution must be applied when applying this model to women who don't receive regular mammographic screening to avoid risk overestimation or other inaccurate results. The model expresses the log of the odds ratio for disease as:

$$\begin{aligned}\log O(D:\overline{D}) &= -0.74948 + 0.09401m + 0.52926b \\ &\quad + 0.21863l + 0.95830n + 0.01081a \\ &\quad - 0.28804b \times a - 0.19081l \times n\end{aligned}$$

where

m = age at **m**enarche

b = number of previous breast **b**iopsies

l = age at first **l**ive birth

n = **n**umber of first degree relatives who have breast cancer

a = 1 if **a**ge≥50, 0 otherwise

This equation suggests that breast cancer risks increase with older age at menarche, more number of previous breast biopsies, older age at first live birth, more number of first degree relatives who have breast cancer, and older age. It is worth noting that the model also compensated for the covariance of two pairs of variables: number of previous breast biopsies and age, and the covariance of age at first live birth and number of first degree relatives who have breast cancer, using the last two terms with negative coefficients. That is, the increased risk of breast cancer due to the two risk factors together in either of the above two pairs is less than the sum of the effects of the two risk factors alone. In other words, there are interactions between the two risk factors in either pair and their effects on breast cancer are dependent.

The equation is used to derive relative risk only, because it is trained on a case-control data set. In order to predict absolute risks, a baseline risk must be established for a specific configuration of the variables, for which the relative risk is 1. The absolute risks of women with different configurations can then be calculated by multiplying this baseline risk with the relative risks

derived from the equation. This absolute risk is then projected to get long-term probabilities . The Gail model has been widely used in breast cancer counseling and research subject screening.

Anderson *et al.* modified the original Gail model, or Gail model 1, to project the risk of developing invasive breast cancer, and this model was referred to as model 2. [28] They used the same model structure and risk factors, but derived the model parameters from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) instead of BCDDP, and included only invasive breast cancer cases.

Pathak and Whittemore fit a biologically motivated breast cancer incidence rate function to data from published case-control studies conducted in different countries at high, moderate and low incidence of breast cancer. [32] The data include 3,925 breast cancer cases and 11,327 controls interviewed in selected hospitals in 1964-1968. The function parameters specify the dependence of age-specific breast cancer incidence rates on age at menarche, age at menopause, occurrence and timing of full-term pregnancies, and body mass. They reported three patterns: "1) Incidence rates jump to a higher level after first childbirth, but then increase with age more slowly thereafter. 2) Rates increase with age more slowly after menopause than before. 3) Rates change quadratically with body mass index among all women, although the main trend varies: Rates decrease with body mass among premenopausal women in high-risk countries, but increase with body mass in all other groups of women." [32]

In the late 90's, Colditz and Rosner developed a log-incidence model of cumulative breast cancer risks to incorporate temporal relations between risk factors and incidence of breast cancer. [34, 35] They evaluated reproductive history, benign breast disease, use of postmenopausal hormone, weight, alcohol intake, and family history as risk factors, and derived the model based on the Nurses' Health Study (NHS). Colditz *et al.* also modified this model to fit incidence data from patients having breast cancer with specific estrogen receptor (ER) and progesterone receptor (PR)

status, and the result suggested better discrimination ability on ER positive and PR positive incidence than on ER negative and PR negative incidence. [36] Part of our research is based on this model, and it will be introduced in more detail later.

2.3 Evaluation and Validation of Breast Cancer Models

After reviewing many breast cancer models from the literature, it appeared that almost all models were developed in the following steps. The scientists designed a mathematical model based on expert knowledge and known risk factors, selected a target population, then fit the model to the data, in many cases using logistic regression, to derive model parameters, relative risks and sometimes absolute risks. The natural question is: how well do these models perform in practice, on general population data?

To answer this question, models need to be evaluated and validated. In machine learning, it is a common practice to build the model on training set data, and then evaluate the model on a separate test set. Methods such as cross-validation and leave-one-out are also popular. In the breast cancer research described above, model validation did not receive as much attention as the models themselves. The Gail model, however, as one of the most widely used models clinically, was validated in a few studies.

Bondy *et al.* evaluated Gail model 1 in 1994 on a cohort of women who participated in the American Cancer Society 1987 Texas Breast Cancer Screening Project and had a family history of breast cancer. [57] They compared the observed (O) and expected (E) breast cancer incidence, and found that Gail model 1 had a better performance among women who adhered to the American Cancer Society mammographic screening guidelines (O/E = 1.12, 95% CI: 0.75-1.61) than it did for those who did not adhere to the guidelines (O/E = 0.41, 95% CI: 0.2 - 0.75). This finding confirmed that the Gail model overestimates breast cancer risks for women not taking annual mammographic screening. They also employed the Hosmer-Lemeshow Goodness-of-fit

test, which did not find an overall lack of fit between the observed and expected number of breast cancers, despite the overestimation result. (When a goodness-of-fit test gives no significant lack of fit result, it does not prove nor guarantee a good fit. Actually this test can only prove lack of fit, but it is always helpful to report a null finding of this test.)

About the same time, Spiegelman *et al.* evaluated Gail model 1 on another cohort of women from the Nurses' Health Study, showing that the model overestimated risk among premenopausal women, women with extensive family history of breast cancer, and women with age at first birth younger than 20 years. [58] They also evaluated the model using the correlation coefficient between observed and predicted risk, which was 0.67.

Costantino *et. al* evaluated both Gail model 1 and Gail model 2 on data from women enrolled in the Breast Cancer Prevention Trial. [38] They compared the ratio of expected to observed number of breast cancers, and the result showed better performance by model 2 than model 1, which underestimated breast cancer risk in women more than 59 years of age.

Rockhill *et al.* evaluated Gail model 2 both on goodness of fit and its discriminatory accuracy using the Nurses' Health Study data. [59] They evaluated goodness of fit by comparing the ratio of expected to observed number of breast cancers, and evaluated discriminatory accuracy using the concordance statistic (i.e. C-index, equivalent to Area Under ROC curve). They also compared the highest and lowest deciles of relative risks derived from Gail model 2 to get a range of discrimination of the model. They reported that the model fit well in the sense of predicting stratified breast cancer incidence, but with modest discriminatory accuracy.

Recently, Gail *et al.* published a paper on criteria for evaluating models of absolute risk, showing that the community is now paying more attention to model risk prediction evaluation. [121] In their paper, Gail summarized general criteria for assessing absolute risk models, including

calibration, discrimination, accuracy, and proportion of variation explained.

For calibration evaluation, they cited goodness-of-fit statistics and comparison of expected number of cases with observed number of cases from [38] and [59], and also mentioned the Brier statistic, or mean squared error, which is a combined measurement of calibration and discrimination. For discrimination evaluation, they reviewed area under the receiver operator characteristics (ROC) curve, which has been widely used, and the Lorenz curve, which is more frequently used in economics research. The Lorenz curve describes the relationship between the proportion of disease cases and the proportion of population that has a risk up to a specific risk value, and hence measures the concentration of risk populations. For rare diseases, the Lorenz measurement is approximately the same as the ROC curve. They also mentioned the measurement of proportion of variation explained using entropy and fractional reduction in entropy.

2.4 Breast Cancer and Genomics

The human genome has roughly 3 billion base pairs (bp), and now is estimated to have 30,000 to 40,000 genes [74]. It has been a well-known fact that most of the base pairs in the genome sequence are identical across the population, while approximately one out of a thousand base pairs will be different when comparing the genome sequences from two different persons. Such differences, or polymorphisms, are used as markers in gene mapping and linkage analysis.

Genetic polymorphism is defined as “the occurrence of multiple alleles at a locus, where at least two alleles appear with frequencies greater than 1 percent,” or a heterozygote frequency of at least 2 percent [81]. Commonly used polymorphisms include restriction fragment length polymorphism (RFLP), variable number of tandem repeats (VNTR), microsatellites, and single nucleotide polymorphisms (SNPs).

A SNP is a sequence polymorphism differing in a single base pair, and is the most common type of polymorphism. SNPs can occur in the gene coding regions, while other types of polymorphisms mainly occur in non-coding regions. Such properties make SNPs the best marker for gene mapping and linkage analysis, because they are common (about 1 per thousand base pairs) and can be very close to the genes and mutations of interest.

SNPs have been reported from many research groups [82-85], while an accumulated public SNPs database (roughly 1.2 million SNPs) can be found on line at several websites, including the ENSEMBL (http://www.ensembl.org), NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov), and TSC (http://snp.cshl.org) websites. The estimated total number of SNPs in the human genome may be over 10 million [86], perhaps as many as 30 million [87].

Breast cancer has long been known to be related to family history, and therefore, to be a hereditary disease. Two relatively high penetrance genes, BRCA1 and BRCA2, have been identified [88-90], but they do not account for all hereditary breast cancers. At least one susceptible area of another major breast cancer gene has been proposed [91-92], while quite a few polymorphisms have been investigated, including rare genetic syndromes associated with increased breast cancer risks and low penetrant breast cancer susceptibility genes. These suspected genes include proto-oncogenes (HRAS1), metabolic pathway genes (NAT1, NAT2, GSTM1, GSTP1, GSTT1, CYP1A1, and CYP1B1), estrogen pathway genes (CYP17 and CYP19), estrogen receptor gene (ER), progesterone receptor gene (PR), androgen receptor gene (AR), and many other genes (COMT, UGT1A1, HLA, TNF[alpha], HSP70, HFE, TFR, VDR, and VPC) [93].

However, convincing results are hard to achieve. Some of the studies show positive association of one polymorphism with breast cancer while others show mild or no association of the same polymorphism. For example, Helzlsouer [94] and Charrier [95] suggested positive association

between GSTM1 and breast cancer risk, while Ambrosone [96] and Garcia-Closas [102] showed evidence against this hypothesis. Another example is that Kristensen [97] reported increased breast cancer risk with CYP19, and Haiman's reports [98-99] find no evidence of positive association. The current findings need to be treated with caution, and further studies are necessary to make a definitive conclusion.

The lack of affirmative results is possibly due to the fact that breast cancer is a complex trait. Only a relatively small portion of breast cancer is hereditary, and a considerable part of these familial cases are involved with BRCA1 and/or BRCA2, whose strong association may shadow possible weak associations with other genes. In addition, there may be multiple genes interacting with each other or interacting with clinical and life-style factors leading to the incidence. All these issues make it difficult to reach a convincing conclusion.

Some of the studies examined the combinations of a few polymorphisms. Bailey *et al.* examined CYP1A1, GSTM1, and GSTT1, reported no significant association with breast cancer risk of these polymorphisms individually or combined [100]. Huang *et al.* examined CYP17, CYP1A1, and COMT, and reported that COMT genotype has a significant association with breast cancer, either individually or combined with CYP17 and CYP1A1, and CYP17 and CYP1A1 play a minor role in the association [101]. Garcia-Closas *et al.* evaluated the association between GSTM1 and GSTT1 gene polymorphisms and breast cancer risk, and provided evidence against a substantially increased risk of breast cancer associated with GSTM1 and/or GSTT1 homozygous gene deletions [102]. These combined investigations are very few in number compared to the abundant number of studies of a single polymorphism.

Some of the studies investigated certain polymorphisms and a few clinical and life-style factors. Ishibe *et al.* evaluated the associations between the CYP1A1 polymorphisms and breast cancer risk, as well as the potential modification of these associations by cigarette smoking, and report a

suggestive increase in breast cancer risk among women who had commenced smoking before the age of 18 and had the CYP1A1-MspI variant genotype compared to nonsmokers who were homozygous wild type for the polymorphism [103]. Millikan et al. examined the effects of smoking and N-acetylation genetics on breast cancer risk, and reported little evidence for modification of smoking effects according to genotype, except among postmenopausal women [104]. Hunter et al. assessed the relation between NAT2 acetylation status and breast cancer risk, and its interaction with smoking, and reported that cigarette smoking was not appreciably associated with breast cancer among either slow or fast NAT2 acetylators [105].

Clinical and life-style factors other than smoking are also investigated in some studies. Gertig *et al.* examined the associations between meat intake, cooking method, NAT2 polymorphism and breast cancer risk, and observed no significant association between meat intake, NAT2, and breast cancer risk, therefore suggesting that heterocyclic amines produced by high-temperature cooking of meat and animal protein may not be a major cause of breast cancer [106]. Hines *et al.* calculated relative risks and confidence intervals to assess breast cancer risk for ADH3 genotype and alcohol consumption level, and suggested that the ADH3 polymorphism modestly influences the response of some plasma hormones to alcohol consumption but is not independently associated with breast cancer risk and does not modify the association between alcohol and breast cancer risk [107]. Haiman *et al.* assessed the association between the A2 allele of CYP17 and breast cancer risk, and observed that the inverse association of late age at menarche with breast cancer may be modified by the CYP17 A2 allele through endogenous hormone levels [108]. Polymorphisms related to breast cancer also have been studied together with family history and ethnic groups, respectively [109-111].

Limited studies examined more polymorphisms and some clinical and life-style factors. For example, Haiman *et al.* studied 10 polymorphisms in 8 genes and two clinical and life-style factors (menopausal status and postmenopausal hormone use), and suggested some interaction

between UGT1A1 genotype and menopausal status [112], which can possibly be modified by postmenopausal hormone use. These previous successful examples imply that we need to put genotypic data, phenotypic data, and clinical and life-style factors together to explain the complex traits.

The Nurses' Health Study provides an abundant collection of clinical and life-style data as well as personal risk factors for breast cancer, and recently a nested case-controlled group of nurses were genotyped for a collection of SNPs in suspected gene areas. Combing these data together, we have a chance to look for gene interaction and clinical and life-style contributors, and explore their relationships from an overall point of view.

2.5 Machine Learning Techniques

Empirically, machine learning techniques are computer algorithms that attempt to find the best of a class of possible models and to tune its parameters so as to best fit a data set of interest according to specified criteria. When the learning process is completed, and if the model's performance is reasonably good, the model and its parameters will reveal, at least to some extent, the intrinsic structure of the data set.

Machine learning techniques can be divided into two major groups: supervised learning and unsupervised learning. Supervised learning will try to relate the variables in the data set to one specific variable, the class variable, and therefore disclose which variable or variable combinations in the data set are best predictors for the class variable. For example, we can put polymorphism and phenotype data together to make a data set, and use this data set to train a model using one of the supervised learning algorithms. If we pick a variable of interest, say, breast cancer as the class variable, the trained model will try to find the best predictors, maybe one or more of the polymorphism or phenotype variables, or maybe a combination of certain variables (which will require extensive work if using traditional statistical methods to obtain).

Unsupervised learning doesn't have a class variable, and it doesn't focus on the associations to any specific variable. Rather, it explores the associations among all the variables and tries to group them according to their own dependencies. If we use the above hypothetical data set to train an unsupervised model, it will try to find the relationships and possible interactions among all the polymorphisms and phenotypes, including breast cancer, but not exclusively.

Both supervised and unsupervised machine learning are non-hypothesis driven processes. When training the model, the algorithm will automatically search the hypothesis space and try to find a best fit. Hence, such techniques can examine a large number of hypotheses in one single run and save researchers' labor.

The limitation of machine learning is computation power and the problem of overfitting. One critical component of a machine learning algorithm is the model. If the model is not flexible enough, the algorithm doesn't have the power to search a large enough hypothesis space, and thus may miss important possible models. If the model is too flexible, the hypothesis space will be so large that it is impractical to search the whole space. In such situations, we need to determine a search strategy to make good use of the available computation power.

In many applications, the available data set has a very limited sample size compared to the number of variables, thus there will be many hypotheses that can fit the data, causing the overfitting problem. Overfitting can be checked by using a separate testing data set to confirm the trained model. Another way to avoid overfitting is to reduce the number of variables, usually referred to as feature selection.

Bayesian Network Induction of Bayesian networks (BN) is one of the most popular machine learning techniques. Bayesian networks can explore the relationships or dependencies among all

variables within a data set, not restricted to pair-wise models of interactions, and therefore can describe and assess complex associations and dependencies of multiple variables. A Bayesian network (BN) is a directed acyclic graph (DAG), in which the nodes represent statistical variables and the links represent the conditional dependencies among the variables. By looking at the network, one can easily tell the underlying relationships of the variables within the data set, which are now represented by the links connecting the variables. A brief introduction to Bayesian networks will follow in the next section.

Clustering Clustering is an unsupervised machine learning technique, which has been widely used in functional genomics. A clustering algorithm tries to group data into clusters based on certain similarity or distance measurements, and to find natural or intrinsic partitions in the data. It may involve finding a hierarchic structure of partitions (a cluster of clusters).

Support Vector Machine Support vector machine (SVM) is a supervised learning algorithm. A support vector machine tries to maximize the discrimination "margin" between the samples with different class labels. Usually only relatively few samples are at the borders or intersections between different classes, and these samples alone will determine the margin, so they are called "support vectors." SVM uses a kernel function (which can be linear, polynomial, or radial) to find the margin and support vectors. Nonlinear kernels permit the representation of complex discrimination boundaries.

Logistic Regression Logistic regression (LR) is one of the most commonly used machine learning algorithms in medical applications. The underlying model for logistic regression is two Gaussian distributions with equal covariance, and LR tries to fit the data to this model using maximum likelihood criteria. An additional advantage of logistic regression, which probably is the reason that it is popular, is that a trained LR model gives weights for each variable in the form of a likelihood ratio, so that the importance of the correlation of each variable with the class

variable is clearly represented.

Classification Tree A classification tree fits the data with a hierarchical tree structure. At each branch point, the algorithm performs a test on one variable to decide on which branch to continue. The leaves are labeled by the values of the class variable. In the learning processes, the algorithm learns what test (on which variable) to perform, and which label to give each leaf, based on a local optimization of information gain. One advantage of the classification tree method is that the tree structure is a convenient representation of knowledge.

Naïve Bayes Naïve Bayes is a relatively simple classifier based on probability and independence assumptions. It assumes all variables except the class variable are conditionally independent given the class variable. Thus one can calculate the joint distribution by multiplying the marginal distributions of every variable using Bayes' rule. Empirically, even though the independence assumptions don't stand, this algorithm works surprisingly well in many applications.

Ensemble Classifiers Above are a few examples of common machine learning algorithms. Many other algorithms have been developed and tried in various applications, such as genetic algorithm, kernel density methods, K-nearest neighbors, etc. In order to get more robust performance, sometimes multiple algorithms or multiple classifiers with the same algorithm are combined to form a ensemble classifier. Ensemble classifiers often have better performance but require more computation power. Examples include stacking and bootstrapping.

Stacking is a combination of classifiers based on different algorithms. Each classifier is trained on the same training set, and gives its own output. The output of these classifiers and the class variable together constitute a meta data set. A further classifier is trained on this meta data set and the output of this final classifier becomes the final output. [64]

Bootstrapping, or bagging, is an ensemble algorithm that can increase stability and reduce variance. Given a training set D of size N. We creates M new training sets also of size N, each uniformly re-sampled from D with replacement. M classifiers are trained on these M training sets, and their output are combined by voting. [65]

Random forest is an example of an ensemble classifier that consists of many classification trees, each generated using a small portion of the variables randomly picked from all variables. Every tree is trained on a bootstrapping of the training set and not pruned. The classification result of the random forest is the vote of all trees. [66] Such voting can be a plurality vote, or can be a normalized sum of the probability outputs of the trees.

2.6 Bayesian Networks

Bayesian Networks (BN) are Directed Acyclic Graphs (DAG) describing dependency structures among variables. [69] A Bayesian Network uses arrows, or links, to depict the dependency relationship between variables, or nodes. An arrow pointing from variable A to variable B means variable B is dependent on variable A, and vice versa. When there is a link pointing from A to B, A is a parent of B, and B is a child of A. Using these graphical symbols, Bayesian Networks visualize conditional independency structures in such a way that the probability of a variable can be fully described by the probability of its parents, and its conditional probability table given the parents.

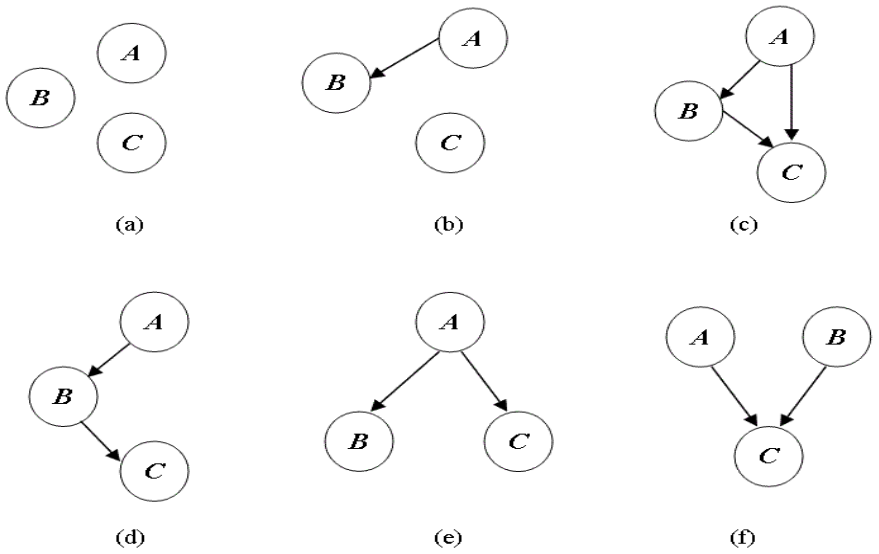

Figure 1 Bayesian Network structures of three variables

Six very simple Bayesian Networks are shown in Figure 1 for illustration. In each of the six figures, there are three nodes representing three random variables A, B, and C. In Figure 1a, the three variables are mutually independent, as completely separated nodes, which corresponds to $P(ABC) = P(A)P(B)P(C)$. In Figure 1b, two variables A and B are dependent, while variable C is independent of A or B. In Figure 1c, all three variables are dependent on each other, corresponding to the general representation of a joint probability distribution $P(ABC) = P(C|AB)P(B|A)P(A)$.

Conditional independence is a very important concept in Bayesian Networks. In Figure 1d, C is dependent on B, which in turn depends on A. A and C becomes conditional independent, or d-separated, given B. The joint distribution can be written as $P(ABC) = P(C|B)P(B|A)P(A)$. In Figure 1e, B and C are d-separated by A and represent a similar conditional independence structure.

Figure 1f shows a different situation where A and B are independent without knowing C, but

becomes dependent given C.

In a Bayesian Network, the set of variables including all parents of variable X, all its children, and all parents of its children is call the Markov blanket for variable X. The Markov blanket will d-separate X from all other variables. In other words, X is conditionally independent of all other variables given the variables in the Markov blanket.

Bayesian Networks have been widely used to represent human expert knowledge and for probabilistic reasoning. In early applications, Bayesian Networks were constructed by asking human experts the dependencies among the variables of interest as well as the conditional probabilities. People have also tried to learn Bayesian Networks directly from data. Current Bayesian techniques to learn the graphical structure of a Bayesian Network from data are based on the evaluation of the posterior probability of network structures [72]. Searching the space of possible network structures has been shown to be NP-hard, [70] but many approximation methods have been developed. We used Bayesware Discoverer as a Bayesian Network learning tool, which applies a greedy search approach. [74][75]

By learning a Bayesian Network from the data, we can obtain a landscape view of the variables and study the interactions among them, as well as an opportunity to discover conditional independency structures that would be overlooked otherwise.

An important difference between a Bayesian Network constructed from human expert knowledge and a Bayesian Network learned from data is that the former often represents causal relationships, while the latter represents conditional independency structures, not necessarily causal or chronological. In statistical and artificial intelligence applications, however, it is a common objective to find causal interpretations of the observed data. Therefore, when learning a Bayesian Network from data with causal interpretations in mind, the variables shall be ordered in such a

way that respects the direction of time and causation. When such ordering is infeasible due to lack of knowledge or other concerns, the learned Bayesian Network must be interpreted with care, and no causal relationship shall be suggested without further investigation.

2.7 Bayesware Discoverer

"Bayesware Discoverer is an automated modeling tool able to transform a database into a Bayesian network, by searching for the most probable model responsible for the observed data." [73] Discoverer represents Bayesian Networks with a graphical interface as nodes and directed edges, and integrates convenient analysis tools to manipulate the network and variables, as well as to display useful information such as conditional distribution tables and Bayes Factors (log likelihood ratios between possible parent sets of a variable). In most of the exploratory analysis of this work, we employed Discoverer as the major tool.

Discoverer can learn both the structure and parameters (or the parameters of a given structure) from the data, based on the K2 algorithm proposed by Cooper *et al.* in 1992. [74][75] As is necessarily typical of computer programs that approximate NP-hard problems by heuristic techniques, Discoverer runs efficiently enough to be useful, but cannot explore the vast space of all possible hypotheses. Therefore, the network it chooses to fit a set of data may not, in fact, be the most likely ("best") network.

Discoverer searches the space of possible conditional dependency structures using a greedy approach based on a list of ordered variables. When searching for the optimal parent set of a variable X, Discoverer uses a greedy approach, i.e. compare the log-likelihood score of a current parent set (start from empty set) and that of the current set plus any additional candidate parent, and update the current set with the additional parent if the score is higher, and then repeat this process until no more parent can be added. The ratios between the log-likelihood score of this final parent set and the other possible parent sets are called Bayes Factors (BFs), and a high BF

generally suggests that the final parent set is much more probable than others. BF's are often astronomically large, because even a slightly better-fitting model can make a large data set enormously more likely. Consequently, a BF that appears large, say 100, may not be very significant.

Discoverer uses the ordering of the variables to guarantee the acyclic constraint in a way such that only variables ordered before variable X are eligible as a candidate parent for variable X. We sometime put the variables in temporal order, hoping that the found dependencies will be consistent in temporal succession and thus could possibly reveal causal relationships, while sometime we order the variables to emphasize certain dependencies based on expert knowledge and known causal relationships. More details will be discussed in later chapters.

Chapter 3 Landscaping Clinical and Life-style and Genotypic Data

In this chapter, we review our attempt to explore the dependencies among a group of clinical and life-style variables, personal risk factors, and a few Single Nucleotide Polymorphisms (SNPs) using Bayesian Networks. The data are drawn from the Nurses' Health Study. We also describe the dependency between breast cancer and clomid use, a possible new risk factor that we found in the exploratory analysis.

3.1 Landscaping Clinical and Life-style Variables and SNPs

Breast cancer has long been known to be related to family history, and therefore, to be a hereditary disease. The recently identified BRCA1 and BRCA2 mutations are responsible for only part of hereditary breast cancer. There could be a "BRCA3," or some modifier genes and clinical and life-style contributors. Most of the previous research on breast cancer has focused on only one or a few genes, and rarely considered the influence of clinical and life-style factors. The lack of definite results from past research illustrates that breast cancer is a complex trait and therefore exploring the susceptible genes and possible contributing clinical and life-style factors systematically can be a fruitful alternative to evaluating the genes individually and separately.

NHS provides an abundant collection of clinical and life-style data as well as personal risk factors for breast cancer, and recently a nested case-control group of nurses were genotyped for a collection of SNPs in suspected gene areas. Combing these data together, we had a chance to look for gene interactions and clinical and life-style contributors, and explore their relationships from an overall point of view.

3.1.1 Data

The genotypic data we had is provided by Ms. Rong Chen in the Channing Lab, including 12 SNPs picked from all SNPs genotyped in the nested case-control study by Dr. David Hunter in the Channing Lab. These SNPs are:

- atm1, atm2, atm3, atm4, and atm5: hyplotype-tagging SNPs in ATM (ataxia telangiectasia mutated protein) gene
- ephx1 and ephx2: non-synonymous SNPs in epoxide hydrolase
- vdrbsm11 and vdrfok1: SNPs that make Restriction Fragment Length Polymorphisms in the vitamin D receptor gene
- xrcc3466, xrcc3471, and xrcc3472: SNPs in XRCC3 (X-ray repair complementing defective repair) gene

The data contain 1007 incident breast cancer cases and 1416 controls. The cases were selected from breast cancer cases diagnosed by June 1, 1998, excluding any other prior cancer diagnosis except for non-melanoma skin cancer. The controls were matched on year of birth, menopausal status, PMH use at time of blood draw (1989), and time of day, time in the menstrual cycle and fasting status at the time of blood draw.

The clinical and life-style data we had include 97 variables, manually selected among the thousands of variables in NHS by Dr. Graham Colditz, Dr. Karen Corsano, and Ms Lisa Li from the Channing Lab (different from the data in the example given at the beginning of this chapter). The variables range from general information, such as race and age, to very specific information, e.g. usage of specific medications. It also includes suspected or known risk factors for breast cancer such as parity and menopausal status. A list of the variables and their meanings can be found in Table 1. These data records are from various years (many from 1990), and most of them don’t include temporal information.

Table 1 Variable list for exploratory analysis

Name	**Meaning**
actmet	Met score of activity measurement
adep	Alcohol dependent problem
agedx	Age of diagnosis of breast cancer
agefb	Age of giving birth to first child
agesmk	Age started smoking
amnp	Age at menopause
asprin	Whether takes Aspirin
ball	Sunburn overall
bback	Sunburn on back
bbd	Benign breast disease
bface	Sunburn on face
bfeed	Whether was breast fed during infancy
blimb	Sunburn on limbs
bmi	Body Mass Index (BMI)
bp	SF-36 pain index
brfd	Breast fed children
brimp	Breast implant
bwt	Birth weight
case_ctrl	Breast cancer
chol	High cholesterol
clomid	Clomid (a fertility drug) use
crown	Crown-crisp scores, a measure of neurotic symptomatology
dadocu	Occupation of father
db	Diabetes
des	DES (a fertility drug) use
dmnp	Menopause status
dtdth	Date of death
dtdx	Date of birth
durmtv	Durartion of taking multi-vitamin
duroc	Duration of oral contraceptive use
durpmh	Duration of post menopausal hormone use
edu	Education level
era	Estrogen receptor status of breast cancer
fhxbr	Family history of breast cancer
frac	Fracture history
hand	Handedness
hbp	High blood pressure

Name	Meaning
hdye	Hair dye
height	Height
hip	Hip measurement
hrt	Heart disease
hsleep	Hours of sleep
husbedu	Education level of husband
id	ID number of the nurse within NHS
insi	Whether breast cancer is insitu
inva	Whether breast cancer is invasive
kidbwt	Birth weight of the heaviest child
kidsun	Sunburn as a kid
kidtan	Suntan as a kid
kpasmk	Passive smoking as a kid
lipst	Lip stick use
live	Living status (alone or with someone)
lungca	Lung cancer
marry	Marriage status
mh	SF-36 mental health index
mnp	Menopausal status
mnty	Menopausal type
moles	Amount of moles
nod	Whether lymph nodes present at breast cancer diagnosis
ocuse	Oral contraceptive use
oprm	Worked in operating room
ovca	Ovarian cancer
packyr	Total pack-year of smoking
parity	Number of children
pasmk	Passive smoking
Pct10	DIAGRAM of body size at age 10
Pct20	DIAGRAM of body size at age 20
Pct30	DIAGRAM of body size at age 30
Pct40	DIAGRAM of body size at age 40
Pct5	DIAGRAM of body size at age 5
pctfa	DIAGRAM of body size of father
pctma	DIAGRAM of body size of mother
pctnow	DIAGRAM of body size in 1988
pmh	Post menopausal hormone use
pra	Progesterone receptor status of breast cancer
race	Race

Name	Meaning
re	SF-36 role-emotional index
regpd	Regularity of period
rp	SF-36 role-physical index
sf	SF-36 social functioning index
shift	Worked on night shift
slpos	Sleeping position
smkst	Smoking status
snore	Whether snore when sleeping
st15	State living at age 15
st30	State living at age 30
state	State currently living
stborn	State born
tagamet	Tagamet (antiacid) use
talcum	Talcum powder use
tamox	Tamoxifen (a drug used in breast cancer treatment and prevention)
tb	Tuberculosis
tubal	Tubal ligation
tv	Time watching TV
vt	SF-36 vitality index
waist	Waist measurement
work	Working status

3.1.2 Exploring the SNPs and Breast Cancer

In the first step, we learned a Bayesian Network from the SNPs data plus a variable marking breast cancer cases or controls. We used all available data, and imputed the missing values using K-nearest neighbor method. This network is shown in Figure 2

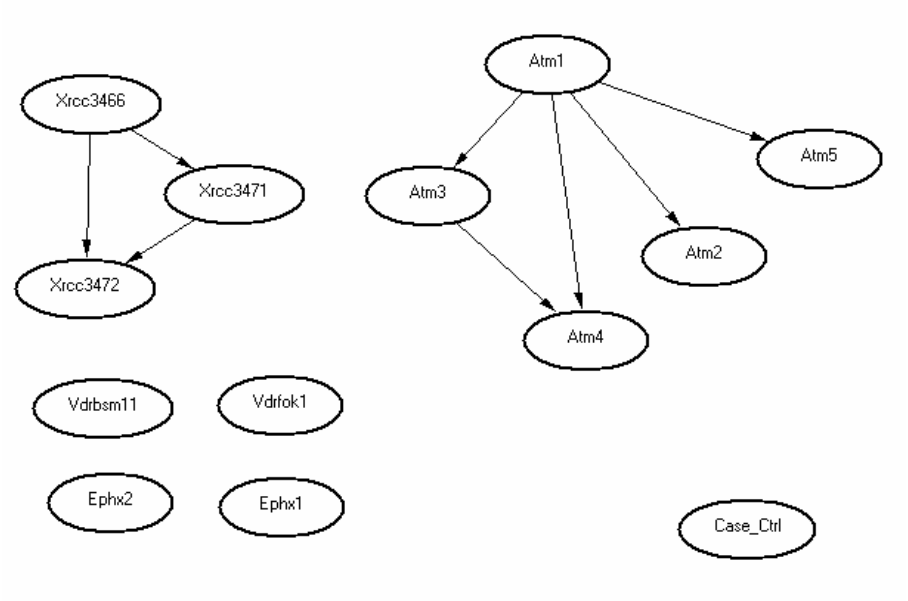

Figure 2 Bayesian Network of 12 SNPs and breast cancer

The Bayesian Network shown in Figure 2 didn't find any dependencies between the SNPs and breast cancer (the BF between breast cancer having no dependent variable and having one dependent variable is 97, suggesting no dependencies with a high confidence), though it did show dependencies among the SNPs from the same genes (atm1-atm5, xrcc3 SNPs). We also tried removing all missing values, but that result didn't show any dependencies between the SNPs and breast cancer, either.

Considering the fact that some breast cancer may not be genetically related, and that breast cancer genes generally have low penetrance, we removed from the data breast cancer cases without a family history of breast cancer and controls with a family history. We hoped to reduce the noise caused by the complexity of breast cancer as a hereditary disease with complex traits, and to increase the chance to find the dependency between the SNPs and breast cancer, if there is any. After such manipulation, we got a sub set with 152 cases and 1274 controls. The Bayesian Network learned from this sub set has the same structure as in Figure 2, showing no dependencies between the SNPs and breast cancer.

While SNPs (or genes) alone are independent of breast cancer, there might be some clinical and life-style factors that can "turn on" these genes. To explore such possible interactions, we need to study the SNPs and the clinical and life-style variables together.

3.1.3 Exploring Clinical and Life-style Factors and SNPs

In the second step, we put the SNPs data together with the clinical and life-style variables to construct a combined data set. This data set contains 96 variables, 12 SNPs and 84 clinical and life-style variables. 13 out of the 97 clinical and life-style variables were not used because of different reasons. Some variables were not included because they were not applicable for this analysis, such as id, and date of death (dtdth). Some variables were combined with other variables to reduce the search space, including pack year (packyr) combined with smoking status, date of diagnosis (dtdx) combined with age of diagnosis, duration of oral contraceptive use (duroc) combined with oral contraceptive use, duration of post menopausal hormone use (durpmh) combined with post menopausal hormone use, and menopause status (mnp) combined with menopause type (mnty). Some variables are characteristics of breast cancer, and thus not included in the Bayesian Network learning, but were used to select cases in a later study. These include estrogen receptor (era), progesterone receptor (pra), in-situ (insi), invasive (inva), and nodular (nod). One variable (kidbwt, birth weight of heaviest child) was removed because of too many missing values.

We used all records and imputed the missing values using K-nearest neighbor method. The variables are ordered in a way such that closely related variables are grouped together, while we tried to make the overall order consistent with temporal order, as in the example we discussed at the beginning of this chapter. The detailed order of variables is listed in Table 2.

Table 2 Variable orders for exploring clinical and life-style factors and SNPs

#	*Variable*	#	*Variable*	#	*Variable*	#	*Variable*
1	atm1	25	marry	49	hdye	73	regpd
2	atm2	26	husbedu	50	moles	74	amnp
3	atm3	27	pasmk	51	kidsun	75	dmnp
4	atm4	28	oprm	52	kidtan	76	mnty
5	atm5	29	shift	53	bback	77	pmh
6	ephx1	30	pctma	54	bface	78	hbp
7	ephx2	31	pctfa	55	blimb	79	chol
8	vdrbsm11	32	pct5	56	ball	80	db
9	vdrfok1	33	pct10	57	asprin	81	tb
10	xrcc3466	34	pct20	58	tagamet	82	hrt
11	xrcc3471	35	pct30	59	talcum	83	frac
12	xrcc3472	36	pct40	60	tamox	84	bbd
13	race	37	pctnow	61	clomid	85	lungca
14	dadocu	38	height	62	des	86	ovca
15	stborn	39	waist	63	tubal	87	actmet
16	st15	40	hip	64	ocuse	88	rp
17	st30	41	bmi	65	parity	89	bp
18	state	42	slpos	66	agefb	90	vt
19	fhxbr	43	snore	67	brfd	91	sf
20	kpasmk	44	lipst	68	adep	92	re
21	bwt	45	brimp	69	work	93	mh
22	bfeed	46	agesmk	70	live	94	crown
23	hand	47	smkst	71	hsleep	95	agedx
24	edu	48	durmtv	72	tv	96	case_ctrl

The Bayesian Network learned from these 96 variables is shown in Figure 3. The majority of the graph is occupied by the clinical and life-style variables, while the SNPs are at the upper right end of the graph, marked by the circle.

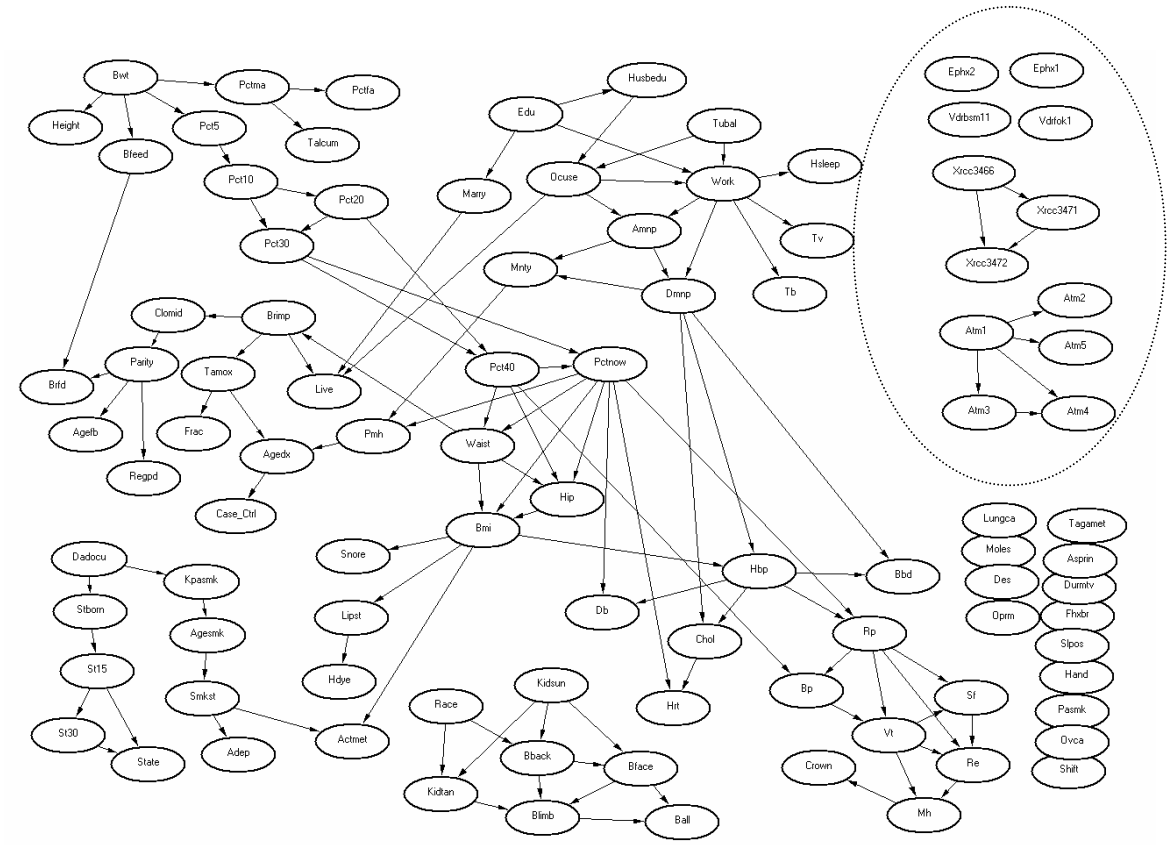

Figure 3 Bayesian Network for interactions among clinical and life-style variables and SNPs

This network presents a landscape of the clinical and life-style variables and describes their interactions. It also shows, again, the dependencies among the SNPs from the same genes. However, this result still fails to show any dependencies between the SNPs and breast cancer, or any other phenotype variables. In later studies, we also tried to limit the data to specific breast cancer type, estrogen receptor positive and progesterone receptor positive, and those results show no dependency between breast cancer or other clinical and life-style variables and the SNPs, either.

Because we were also looking for any variable whose presence would make breast cancer and any SNPs become dependent, we tried different variable orders to facilitate such searching as well (e.g. putting breast cancer and SNPs on top of all other variables), but did not find any such dependencies, either.

In our attempt to explore the relationship between breast cancer and a small collection of SNPs from different genes, we didn't find any dependency that links a SNP and any clinical and life-style variable, including breast cancer. There are a few possible reasons. It could be that the selected SNPs are truly not associated with breast cancer, as suggested by the null finding with a reasonable amount of effort to look for dependencies. It could also be because one or more SNPs in our study are associated with a mutant gene that does contribute to breast cancer, but only triggered by certain clinical and life-style factors that are not included in our study. Considering the wide range of variables included in this study, this is not very likely, and even if it is true, the association must be weak or the triggering clinical and life-style factor must be rare in daily life, or we would have found a link between that SNP and breast cancer, at least. Another possibility is that the association between SNPs and breast cancer is very weak, and covered by the random noise.

Considering the size of the study, both in terms of the number of cases and variables, we consider the first assumption above, that the selected SNPs are not associated with breast cancer, a reasonably plausible result of this work.

3.2 Estrogen Receptor Status and Clomid

During the exploration process in landscaping the clinical and life-style variables, we noticed that in many of the Bayesian Networks we learned, there is a dependency between history of using clomid (a fertility drug) and estrogen receptor status of breast cancer, as shown in Figure 4. (Only part of the network is shown for a clear look.) This network is built on the data set that we used for landscaping, thus matched on age and menopausal status for breast cancer. After removing records with missing values on more than 12 variables, the data set used to learn this Bayesian Network contains 514 ER+ breast cancer nurses, 151 ER- breast cancer, and 834 non-breast cancer.

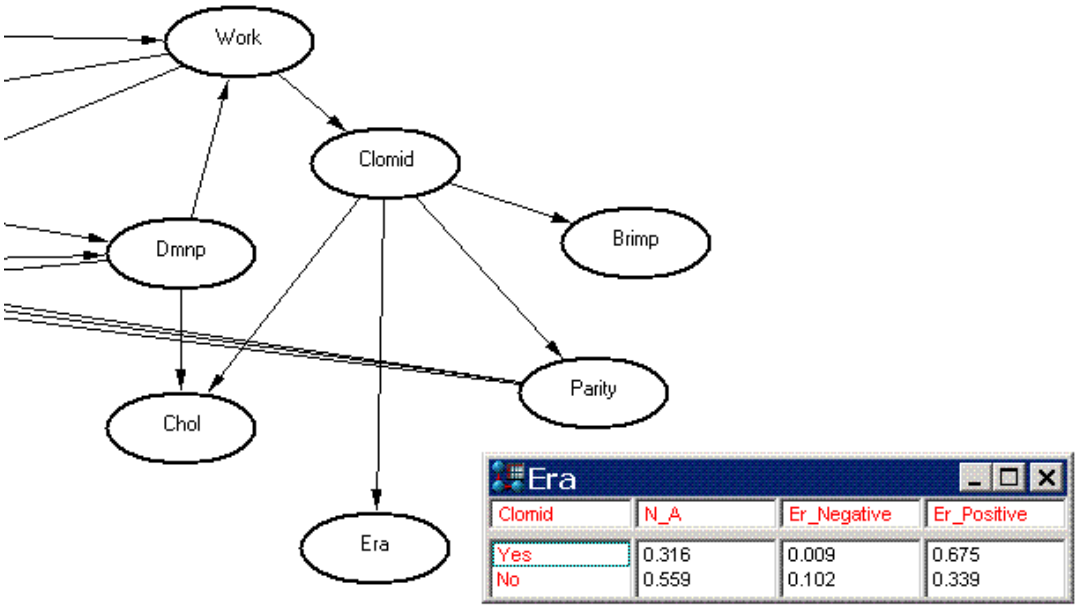

Figure 4 Clomid use and breast cancer with ER status

In the graph, the conditional distribution table of ER status of breast cancer (Era) is also shown. From this table, we see that clomid use is associated with lower ER- breast cancer incidence (0.316 vs. 0.559) and higher ER+ breast cancer risk (0.675 vs. 0.339).

Other variables shown in the graph are the Markov blanket of clomid, which d-separates clomid from the rest of the graph. Among these variables, clomid use is positively associated with breast implant (Brimp) and negatively associated with number of children (Parity), high cholesterol (Chol), post menopause (Dmnp), and retirement (Work).

The association with number of children can be easily understood: nurses with more children are less likely to take fertility drugs. Noting the fact that the data are age matched for breast cancer, not clomid, most other associations could probably be explained by the greater use of clomid in a younger generation of nurses than among older ones, because when clomid became clinically

available in 1968, [71] older nurses in NHS were already past their reproductive age. Figure 5 below illustrates this distribution over age for nurses who took or never took clomid. In the graph, we can see the birth years of nurses who didn't take clomid were approximately evenly distributed from 1921 to 1946, while most of the nurses who took clomid were born after the mid 30's.

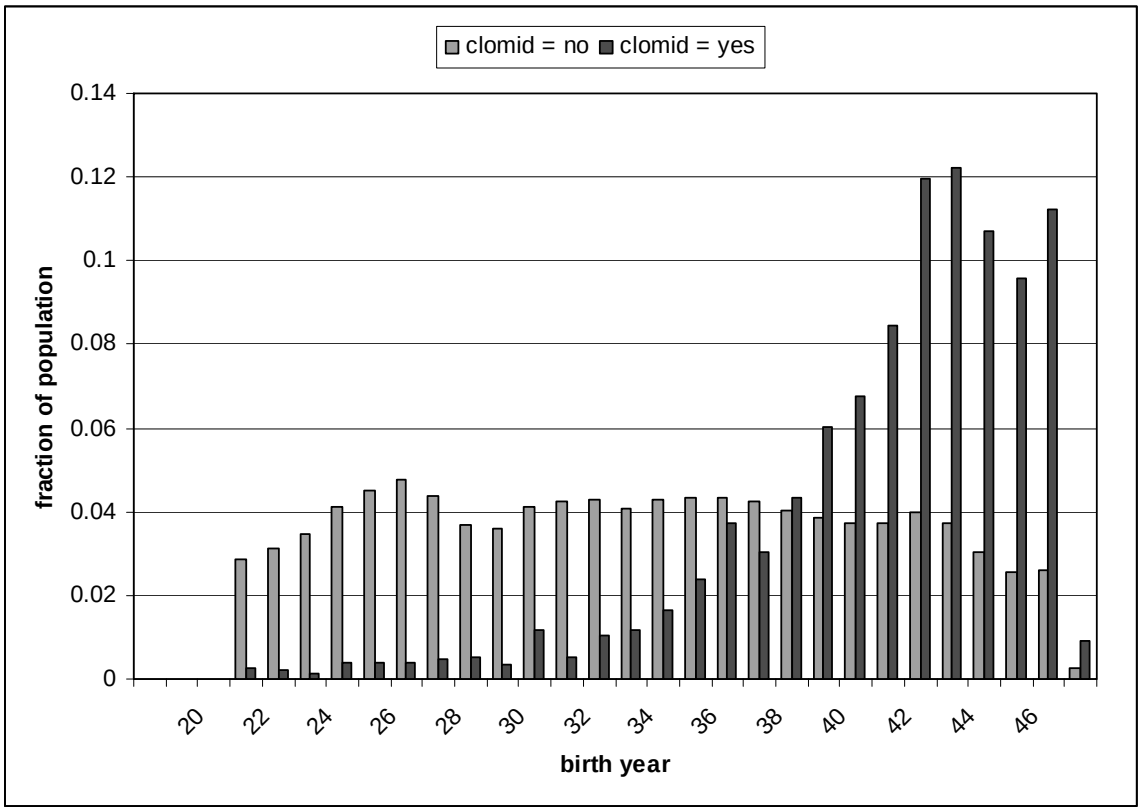

Figure 5 Comparison of birth year between nurses who used and never used clomid

It is also worth noting that the clomid data were collected in 1992 and 1994, when most of the nurses in NHS passed their reproductive age, so it is unlikely that any nurses would start taking clomid after that data collection point, and therefore, there should be no bias introduced by censoring effects on clomid data.

For the dependency between clomid use and ER status of breast cancer, it cannot be explained by

age or generation difference, because the data were age-matched for breast cancer. To further investigate this dependency, we used the full cohort data, excluding those missing clomid or ER status.

Before looking at the statistics, we first checked the date of taking clomid and the date of diagnosis for breast cancer, and found that no clomid was taken after the date of diagnosis for breast cancer, while 10 nurses reported clomid use without specific dates, and 5 nurses who took clomid at age 24, 25, 32, 34, 36 were diagnosed with breast cancer but without date of diagnosis. These nurses were included in the study.

We compared clomid use versus ER status of breast cancer from the full cohort (excluding missing values on clomid use and ER status), in Table 3. The table is organized in three groups, showing pair-wise comparison of the effect of clomid use on ER- breast cancer (ER-), ER+ breast cancer (ER+), and non breast cancer (non-brcn). In all three groups, column 2 through 4 shows the number of nurses in each category. Column 4 is the χ^2 value based on the null-hypothesis that the compared pair (two of ER+, ER-, or non-brcn, indicated in the head row of each group) and clomid are independent, and column 5 is the corresponding p-value, the probability that the null hypothesis is true, or that clomid use is associated with the compared pair. Column 6 is the odds ratio that the odds of the compared pair given clomid=y divided by the odds given clomid=n, therefore reflecting the influence of clomid use on odds of the two variables.

The data in Table 3 shows that clomid use is associated with lower estrogen receptor negative (ER-) breast cancer, indicated by odds ratio of 0.47, with no significant impact on estrogen receptor positive (ER+) breast cancer.

Table 3 Clomid and breast cancer with ER status, no matching

# cases	ER-	ER+	Sum	Chi-square	p-value	Odds ratio	95% CI for Odds ratio
clomid=y	8	60	68				
clomid=n	953	3245	4198				
Sum	961	3305	4266	4.586	0.032	0.45	0.22-0.95
	ER-	non-brcn	Sum				
clomid=y	8	1437	1445				
clomid=n	953	80358	81311				
Sum	961	81795	82756	4.731	0.030	0.47	0.23-0.94
	ER+	non-brcn	Sum				
clomid=y	60	1437	1497				
clomid=n	3245	80358	83603				
Sum	3305	81795	85100	0.063	0.80	1.03	0.80-1.34

As discussed above, clomid use is dependent on age or generation of the nurses. Therefore, we matched clomid use on age. For every nurse with clomid usage, we randomly picked 13 nurses who never used clomid. Because the minimum ratio of nurses who never used clomid to nurses who used clomid at different ages is 13, we could find at most 13 nurses with a history of non clomid use to match every nurse who used clomid. These age-matched data are shown in Table 4.

Table 4 Clomid and breast cancer with ER status, matched on age for clomid use

# cases	ER-	ER+	Sum	Chi-square	p-value	Odds ratio	95% CI for Odds ratio
clomid=y	8	60	68				
clomid=n	200	520	720				
Sum	208	580	788	8.200	0.0041	0.35	0.16-0.74
	ER-	non-brcn	Sum				
clomid=y	8	1437	1445				
clomid=n	200	18729	18929				
Sum	208	20166	20374	3.361	0.067	0.52	0.26-1.06
	ER+	non-brcn	Sum				
clomid=y	60	1437	1497				
clomid=n	520	18729	19249				
Sum	580	20166	20746	8.725	0.0031	1.50	1.15-1.98

In the age-matched data, we see a slightly weaker negative association of clomid with ER- breast

cancer, and a new positive association of clomid with ER+ breast cancer. This change in the odds ratios was as expected because it is known that ER+ breast cancer happens more often in older women than in younger women compared to ER- breast cancer, while clomid use happened more often in younger nurses and more older nurses with no clomid use were removed in the matching process.

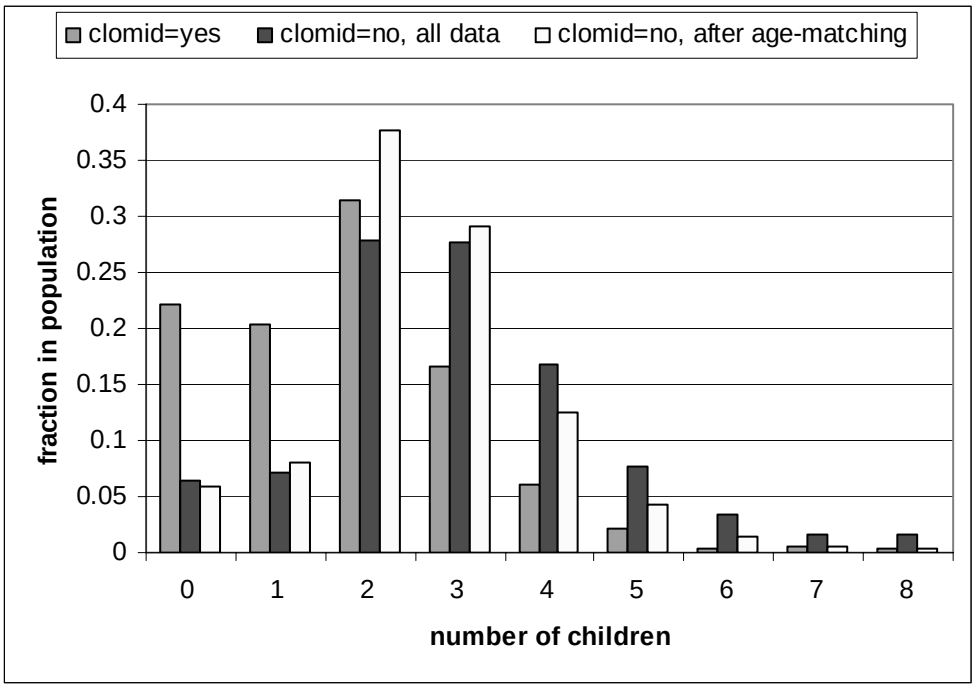

Figure 6 Comparison of number of children on clomid use

However, because the number of children is known to be a risk factor for breast cancer, or specifically, nulliparous is associated with higher risk of breast cancer, we need to check parity of these data as well. The data showed, as in Figure 6, that nurses who took clomid are more likely to have fewer children than those who did not take clomid, which is consistent with our understanding of the purpose of clomid use. Therefore, we matched the data in Table 4 additionally on parity, with a ratio of 3, the maximum number of available matches, and the result is shown in Table 5.

Table 5 Clomid and breast cancer with ER status, matched on age and parity for clomid use

# cases	ER-	ER+	Sum	Chi-square	p-value	Odds ratio	95% CI for Odds ratio
clomid=y	8	60	68				
clomid=n	52	118	170				
Sum	60	178	238	9.128	0.0025	0.30	0.14-0.68
	ER-	non-brcn	Sum				
clomid=y	8	1437	1445				
clomid=n	52	4300	4352				
Sum	60	5737	5797	4.354	0.037	0.46	0.22-0.97
	ER+	non-brcn	Sum				
clomid=y	60	1437	1497				
clomid=n	118	4300	4418				
Sum	178	5737	5915	6.849	0.0088	1.52	1.11-2.09

The statistics shown in Table 5, matched on both age and parity, suggest that clomid use is associated with lower ER- breast cancer incidence versus higher ER+ breast cancer incidence (odds ratio 0.30, 95% CI 0.14-0.68). This effect of clomid on breast cancer can also be interpreted as increased ER+ risk (odds ratio 1.52, 95% CI 1.11 – 2.09), and decreased ER- risk (odds ratio 0.46, 95% CI 0.22 – 0.97).

3.3 Summary

In this chapter, we first described the results from our efforts in exploring data sets from the Nurses' Health Study, including a small selection of SNPs. Even though the results showed dependencies between environment variables and between SNPs from the same gene, we did not find any dependencies between the SNPs and breast cancer, with or without the presence of other clinical and life-style variables as possible "turn-on" factors. Our result suggests that there might be no association between breast cancer and the selected SNPs.

During the exploratory analysis process, we discovered a dependency between clomid use and breast cancer of specific estrogen receptor status in the nested case-control set. In the literature,

different results of studies on clomid (or clomiphene) and breast cancer have been reported. Rossing *et al.* observed that use of clomid as treatment for infertility is associated with lower breast cancer risk (adjusted relative risk: 0.5; 95% CI: 0.2-1.2). This cohort study contains 3837 women, among which 27 are breast cancer cases. [116]

Venn *et al.* reported a transient increase in the risk of having breast cancer in the first year after in-vitro-fertilization (IVF) treatment (relative risk: 1.96; 95% CI: 1.22-3.15). Clomid is one of the fertility drugs used in the study from ten Australian IVF clinics (143 breast cancers, among which 87 were exposed to clomid). [117]

Brinton *et al.* recently reported that clomid use is associated with slight or non-significant elevation of breast cancer risk (adjusted relative risk: 1.39; 95% CI: 0.9-2.1) for subjects followed for more than 20 years, but when restricting the study to invasive breast cancer, the association with clomid was significant (adjust relative risk: 1.60; 95% CI: 1.0-2.5). [118] This study includes 213 (175 for invasive) breast cancer cases, among which 29 (27 for invasive breast cancer) were exposed to clomid and were followed for more than twenty years. [118]

In our study, we found dependencies between clomid use and breast cancer of specific estrogen receptor status, and confirmed these dependencies on the full cohort, checked on the temporal relationship between the time of taking clomid and the time of diagnosis of breast cancer. We also matched nurses who used or never used clomid on age and number of children, and the final result suggests that clomid use is associated with increased ER+ breast cancer risk, and at the same time, decreased ER- breast cancer risk.

Chapter 4 Breast Cancer Model

4.1 Log-Incidence Model for Breast Cancer

Colditz and Rosner developed a log-incidence model of breast cancer to incorporate temporal relations between risk factors and incidence of breast cancer. [34, 35] They also modified this model to fit against breast cancer with specific estrogen receptor (ER) and progesterone receptor (PR) status, and the result showed better discrimination ability on ER positive (ER+) and PR positive (PR+) incidence than on ER negative (ER-) and PR negative (ER-) incidence. [36]

In their paper, Colditz *et al.* made an assumption that breast cancer incidence rate is proportional to the number of breast cell divisions accumulated through a life time. In addition, they also assumed that the rate of increase of breast cell divisions at a certain age is exponential in a linear combination of risk factors. Based on these two assumptions, they calculated the log incidence rate with a linear function of risk factors, which varies with age and change of reproductive life status. This formula is shown below.

$$Log\ I_t = \alpha + \beta_0 (t^* - t_0) + \beta_1 b + \beta_2 (t_1 - t_0)\ b_{1,t-1} + \gamma_1 (t - t_m)\ m_A + \gamma_2 (t - t_m)\ m_B$$
$$+ \delta_1\ dur\_PMH_A + \delta_2\ dur\_PMH_B + \delta_3\ dur\_PMH_C + \delta_4\ PMH_{cur,t} + (\delta 4 + \delta_5)\ PMH_{past,t}$$
$$+ \beta_3\ BMI_1 + \beta_3\ BMI_2 + \beta_4\ h_2 + \beta_4\ h_2$$
$$+ \beta_3\ ALC_1 + \beta_5\ ALC_2 + \beta_5\ ALC_3$$
$$+ \alpha_1\ BBD + \alpha_2\ BBDt_0 + \alpha_3\ BBD\ (t^* - t_0) + \alpha_4\ BBD\ (t - t_m) + \theta\ FHX$$

Where I_t is the incidence rate at age t, and

t = age

t_0 = age at menarche;

t_m = age at menopause;

t^* = min(age, age at menarche);

m_i = 1 if postmenopausal at age I, = otherwise;

s = parity;

t_i = age at ith birth, o = 1,…,s;

b = birth index = $\sum (t^* - t_i) b_{it}$

b_{it} = 1 if parity>i at age t, =0 otherwise;

m_a = 1 natual menopause, =0 otherwise;

m_b = 1 if bilateral oophorectomy, =0 otherwise;

BBD = 1 if benign breast disease = yes, =0 otherwise;

FHX = 1 if family history of breast cancer = yes, =0 otherwise;

dur_PMH_A = number of years on oral estrogen;

dur_PMH_B = number of years on oral estrogen and progestin;

dur_PMH_C = number of years on other postmenopausal hormones;

$PMH_{cur,t}$=1 if current user of postmenopausal hormones at age t, = 0 otherwise;

$PMH_{past,t}$=1 if past user of postmenopausal hormones at age t, = 0 otherwise;

BMI_j = body mass index at age j (kg/m^2);

ALC_j = alcohol consumption (grams) at age j

h = height (inches).

Colditz *et al.* fitted the model with the SAS function PROC NLIN on a selected cohort data set from NHS, and calculated a risk score based on the model, which can be used as an index for breast cancer incidence risks. The data were selected from the NHS cohort by excluding nurses with pre-existing cancers, or missing (or conflict) pregnancy and parity information, or without a precise age at menopause, or missing height and weight information.

In our work, we used a Bayesian approach to explore the same data set, using the same variables in the Colditz-Rosner log-incidence model, to predict breast cancer risk based on incidence rate. Here, breast cancer risk refers to the probability that a woman who doesn't have breast cancer will be diagnosed with breast cancer within a defined period of time. We first built models to

predict the probability that a woman will be diagnosed of breast cancer in two years, then we used similar approach to build models to predict the probability that a woman will be diagnosed of breast cancer in longer terms: four years and six years. Such a breast cancer risk is consistent with the definition of absolute risk of breast cancer that is generally used in this area, [124,121] and it also matches with the probabilities of developing breast cancer predicted in the Gail model. [27] Specifically, Gail *et al.* defined the absolute risk of a disease of interest within a defined period from time *a* to time *t* as

$$\pi \equiv \int_a^t h(u) e^{-\int_a^u [h(v)+k(v)]dv} du,$$

where *h(t)* is the cause-specific hazard at time *t* for the disease of interest, and *k(t)* is the hazard of mortality from other causes at the same time. [121] It is worth noting that such definition also incorporated competing diseases represented by *k(t)*.

4.2 Evaluating Risk Score as an Index for Breast Cancer

Risk scores are widely used as indicators for risks of various diseases, and many of them were calculated using regression models, such as the breast cancer risk score developed by Colditz *et al.*, as described previously. It would be interesting to compare these scores with probabilistic models, and see how they perform differently (or the same).Before starting on the prediction models, we introduce an initial attempt using Bayesian Network to evaluate the risk score as an index for breast cancer developed by Colditz *et al.*, as described previously.

The objective of this analysis is to see whether breast cancer incidence is dependent on the risk score, and whether the risk score can d-separate breast cancer from the risk factors. The underlining assumption is quite intuitive: if the risk score is a good index for breast cancer, it shall capture all information provided by the risk factors and act as an information flow proxy from the risk factors to breast cancer. If there are additional information flow from the risk factors

to breast cancer, not included by the score, it will suggest a possibility of improvement for the risk score by adjusting model and incorporating such information.

Ms. Marion Mcphee in Harvard Medical School provided us a risk score file containing risk scores calculated based on the log-incidence model using 1980 year data. She also provided us a data set including the variables and records that were used to develop the log-incidence model, from 1980 till 1998, which we used to build breast cancer classifiers as described later in this chapter. For the study in this section, we used only 1980 data from this data file and combine with the risk score for analysis, because the risk scores are calculated using 1980 data only, even though the model is developed based on all year data. (When developing risk prediction models in the rest of this chapter, we used all year data.)

In order to learn a Bayesian Network for our purpose, we put breast cancer (Case) on top, risk score (Score_df53) in the middle, and then the risk factors. We hoped that with such a searching order, Discoverer could find a dependency structure that can show whether there is information flow from the risk factors to breast cancer outside the risk score. There were no missing value in this data set, and the continuous variables were discretized into three bins with roughly 1/5 records in the top and bottom bins and 3/5 records in the middle. The result is shown in Figure 7.

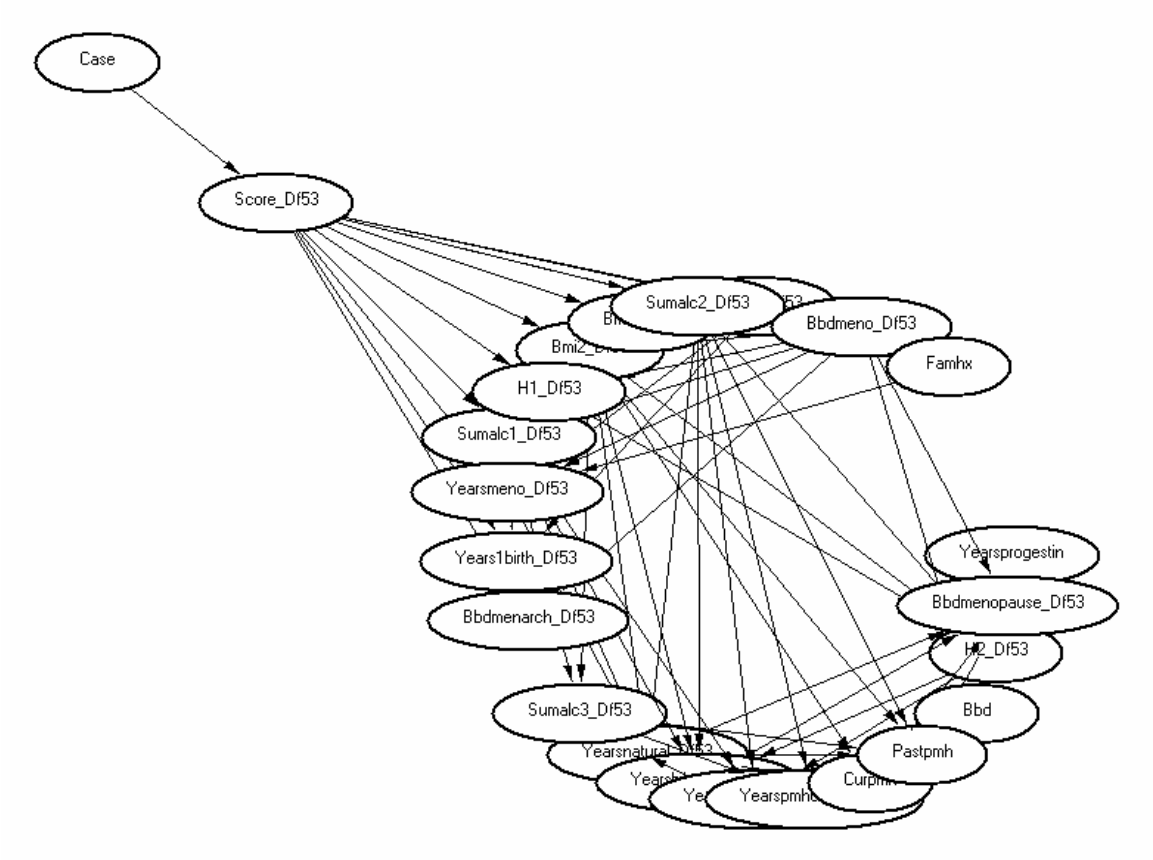

Figure 7 Bayesian Network learned for evaluating risk score as an index for breast cancer

From the learned Bayesian Network, we see that breast cancer is dependent on risk score, which in turns has dependencies with many of the risk factors. There is no direct links between the risk factors and breast cancer, so this result does not show any extra information that the risk score needs to modify to incorporate.

The above result, however, does not guarantee that the risk score captures all information of the risk factors because of two major reasons. First, searching for optimal dependency structure of a Bayesian Network is a NP-hard problem, and the searching method we used does not guarantee a global optimal solution. There could be other dependency structures that are more likely than this structure, and breast cancer and other risk factors might not be d-separated by the risk score in some of those structures. The second reason is that this data set is only one tenth of the whole data set on which the risk score model is developed, and therefore such an evaluation is not

complete. If we had had risk scores from all year data, we would have been able to do a more complete evaluation, while still constraint by the first reason described.

4.3 Learning a Classifier for Breast Cancer Incidence

The original data set we had includes records of 3216 nurses who were diagnosed with breast cancer between 1980 and 2000, and 71805 nurses who were never diagnosed with breast cancer until 2000. NHS data are organized in two-year intervals, because the data were collected every other year. For nurses without breast cancer, there is one record every other year from 1980 until 1998, except for nurses who left the study early due to death, diagnosis of cancer other than breast cancer, or other reasons. These nurses only have records till when they left. For nurses with breast cancer, there is one record every other year from 1980 till the last record before the report of breast cancer.

We used a randomly sampled subset from the above data to train a classifier. We didn't use the whole data set for the following reasons. First of all, the number of records is different for every nurse. Healthier nurses stayed in the study till 2000, thus having more records. Nurses who died or were diagnosed with cancers before 2000, including breast cancer, have fewer records. If we use all available data, healthier nurses will be overemphasized compared with the nurses who exhibited diseases (not limited to breast cancer, but also including other cancers or other diseases that caused mortality). Secondly, using multiple records from one person to train a single classifier might increase the apparent power more than the data can actually provide. Thirdly, for an incidence model, only one record of a nurse with breast cancer will be used as a breast cancer case, and which year to use will depend on the predictive period of the model (discussed in more details later). If we were to use all available data, the proportion of breast cancer cases would be very small (about 0.6%), which is relatively harder to deal with in training classifiers. Lastly, the whole data set is very large (60MB), and sub-sampling it makes the application of the tools we used more practical.

Based on the above reasons, we sub-sampled the data set by randomly picking one record for nurses without breast cancer, while picking the record that immediately precedes the report of diagnosis for nurses with breast cancer (this record is marked as breast cancer). It is worth noting that the sampling ratio used here is different for breast cancer and non-breast cancer, because the sampling of breast cancer is deterministic rather than random. Ideally we should also randomly sample the breast cancer nurses, but in order to make use of all breast cancer cases, we fixed our sampling on the breast cancer record. Therefore, breast cancer cases were amplified, and the risk prediction results of a classifier trained from this data set need to be prorated.

Because Discoverer works better on categorical variables than continuous variables, we discretized continuous variables into three bins, with the two extreme bins containing approximately one fifth of the records each, and the middle bin containing the remaining three fifths. (In the exploratory process, we also tried different discretization methods, including equal-size binning, equal-range binning, discretization on class, discretization on entropy, etc., but did not find better results on performance.)

We randomly split this data set into two parts, 75% as training set, and the remaining 25% as test set. Because breast cancer is a rare disease, and even in this "amplified" data set breast cancer only constitutes about 2% of the records, a little noise in the random splitting may cause noticeable difference in the proportions of breast cancer records in the split sets. Therefore, we stratified the random splitting on breast cancer. Such stratification is necessary to eliminate systematic overestimation or underestimation of risks due to different marginal distributions of breast cancer in training and test sets, which we encountered in preliminary studies.

When learning the network, we organized the variable order in three steps. In step one, we put the variables into an order such that closely related variables are together, while the order is also

consistent with our understanding of the causal relationships among the variables, except for breast cancer. The breast cancer variable, "case," is on top of the variable list, so Discoverer could explore the dependencies between breast cancer and all other variables. (If we put breast cancer at the bottom of the list, Discoverer may not be able to find as many dependencies because of the competition between other variables.)

In step two, we randomly permuted the variable orders to create ten different training sets, with the same data but different variable orders, and learned a group of ten different networks from each training set. For each variable, we counted the number of times it appeared in the Markov Blanket of breast cancer in these ten networks, as in Table 6. From each group, we picked the variable with the highest count (first variables in each group) and moved these variables to the top of the list, to facilitate the searching of the dependencies between these variables and breast cancer.

Table 6 Variable counts in Markov Blankets of breast cancer in ten networks with permuted order

Variable Name	Variable Meaning	Count	Group
age	Age	3	Aging
yearsMeno	Years with menstrual period	2	
yearsNatural	Years after natural menopause	1	
bbdMeno	Years with menstrual period if have benign breast disease	4	Benign breast disease
bbdMenopause	Years after menopause if have benign breast disease	3	
bbd	Benign breast disease	3	
yearsProgestin	Years on progestin	5	Postmenopausal hormone
curPMH	Current postmenopausal hormone use	2	
tmtbm	Birth index	1	Pregnancy
sumalc2	Alcohol consumption when on postmenopausal hormone	1	Alcohol consumption

It is worth noting that family history is not included in these variables. This is probably due to the fact that family history is a very unbalanced variable (only about 10% nurses with family history),

thus being at a disadvantage when competing with other more balanced variables. It is also possible that in all of the limited permutations, family history happened to be in a position that is hard for Discoverer to find the dependency between family history and breast cancer. Nonetheless, it is well known that family history is a predicting factor for breast cancer, supported by the genetic mutations found to cause breast cancer. Therefore, we included family history in the next step of the ordering process despite its absence in Table 6.

In step three, we combined the results from step one and step two, by putting the selected variables in step two, i.e. age, family history, bbdMeno, yearsProgestin, tmtbm, sumalc2 on the top of the list, and obtained the variable order in Table 7.

Table 7 Variable order

Variable Name	**Variable Meaning**
case	Breast cancer
age	Age
famhx	Family history
bbdMeno	Years with menstrual period if have benign breast disease
yearsProgestin	Years on progestin
tmtbm	Birth index
sumalc2	Alcohol consumption when on postmenopausal hormone (PMH)
bmi1	Body Mass Index (BMI) factor without PMH use
bmi2	BMI factor with PMH use
h1	Height factor with PMH use
h2	Height factor without PMH use
sumalc1	Alcohol consumption
sumalc3	Alcohol consumption without PMH use
yearsMeno	Years with menstrual period
years1birth	Years after first birth
yearsNatural	Years after natural menopause
yearsBilateral	Years after bilateral oophorectomy
yearsEstrogen	Years on oral estrogen
yearsPMHother	Years on other PMH than estrogen or progestin
curPMH	Current PMH use
pastPMH	Past PMH use
bbd	Benign breast disease

Variable Name	Variable Meaning
bbdMenarch	Age at menarche if have benign breast disease
bbdMenopause	Years after menopause if have benign breast disease

We learned a Bayesian Network from the training set with the above variable order, shown in Figure 8.

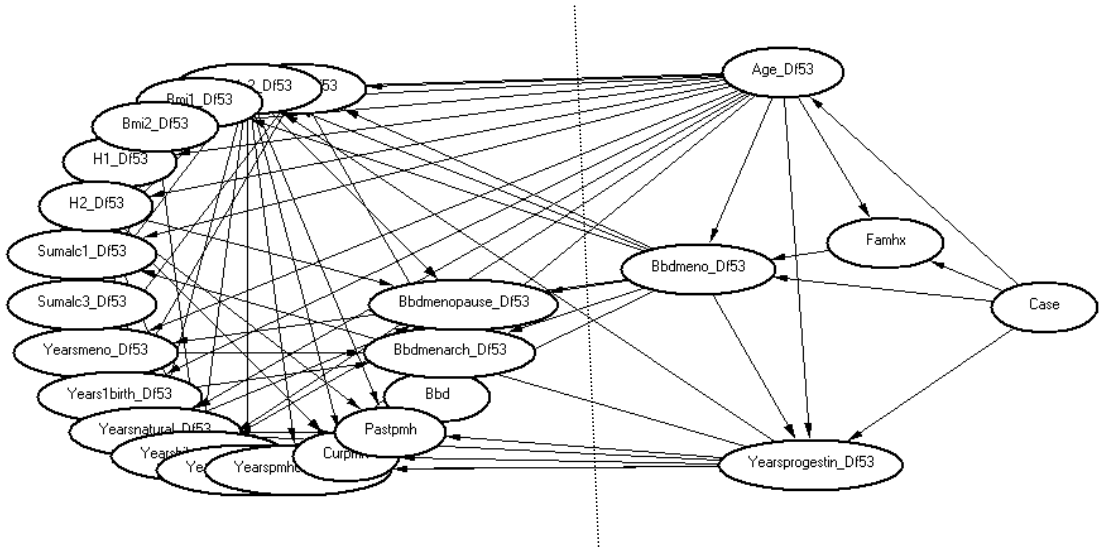

Figure 8 Bayesian Network for predicting breast cancer

In the graph, variable "Case" is the marker for breast cancer. We can see that age (Age_df53), family history of breast cancer (Famhx), years having menstrual period if had benign breast disease (Bbdmeno), and number of years on oral estrogen and progestin (Yearsprogestin_df53) are the four most important predicting factors and constitute the Markov blanket for breast cancer (Case).

We evaluated the generalization error of this model on the hold-out 25% test set. The prediction AUC is 0.65, with 95% Confidence Interval (CI) of 0.64 – 0.67.

4.4 Classifier for ER+/PR+ Breast Cancer

Predicting ER+/PR+ breast cancer is clinically more important, because ER+/PR+ tumors are more sensitive to Selective Estrogen Receptor Modulators (SERMs), one of the major methods for prevention of breast cancer. [125] In further study of breast cancer predicting models, we limited the scope of our model to ER positive and PR positive breast cancer versus non-breast cancer by removing breast cancer case records not marked as ER+/PR+ (including ER-, PR-, and ER or PR missing) from the data set described above, which left 1447 breast cancer nurses and 71805 non breast cancer nurses.

We split (randomly, stratified) the data into a training set with 75% of the total data, and a test set with the remaining 25% of the total data. We built the classifier on the training set, and then ran the classifier through the test set to evaluate its performance, which will be discussed in a later section. The resulting Bayesian Network is shown in Figure 9.

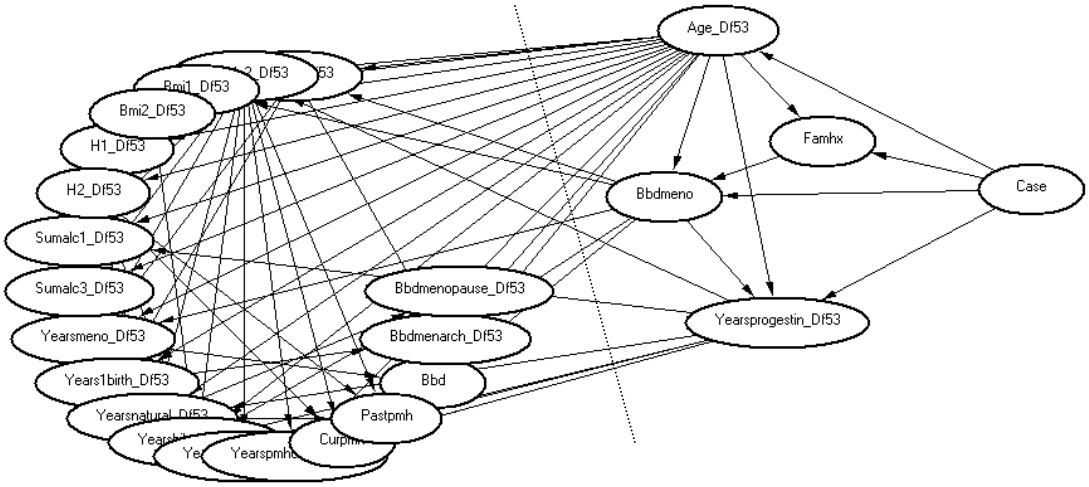

Figure 9 Bayesian Network for predicting ER+/PR+ breast cancer

We can see that the structure of this prediction model is similar to the one in Figure 8 for all breast cancer, with the same predicting variables: age (Age_Df53), family history of breast cancer (Famhx), years of having menstrual period if had benign breast disease (Bbdmeno), and years

taking oral estrogen and progestin (Yearsprogestin_Df53). In the next section, we evaluated this model in more detail.

4.5 Evaluation of Risk Prediction Model

We evaluated the performance of the risk prediction model in two ways. First, we evaluated the discriminating ability of the model using the Receiver Operating Characteristic (ROC) curve, specifically, area under the ROC curve (AUC). Second, we evaluated the model's ability to give an accurate probability for breast cancer prediction, by comparing expected and observed incidence rates.

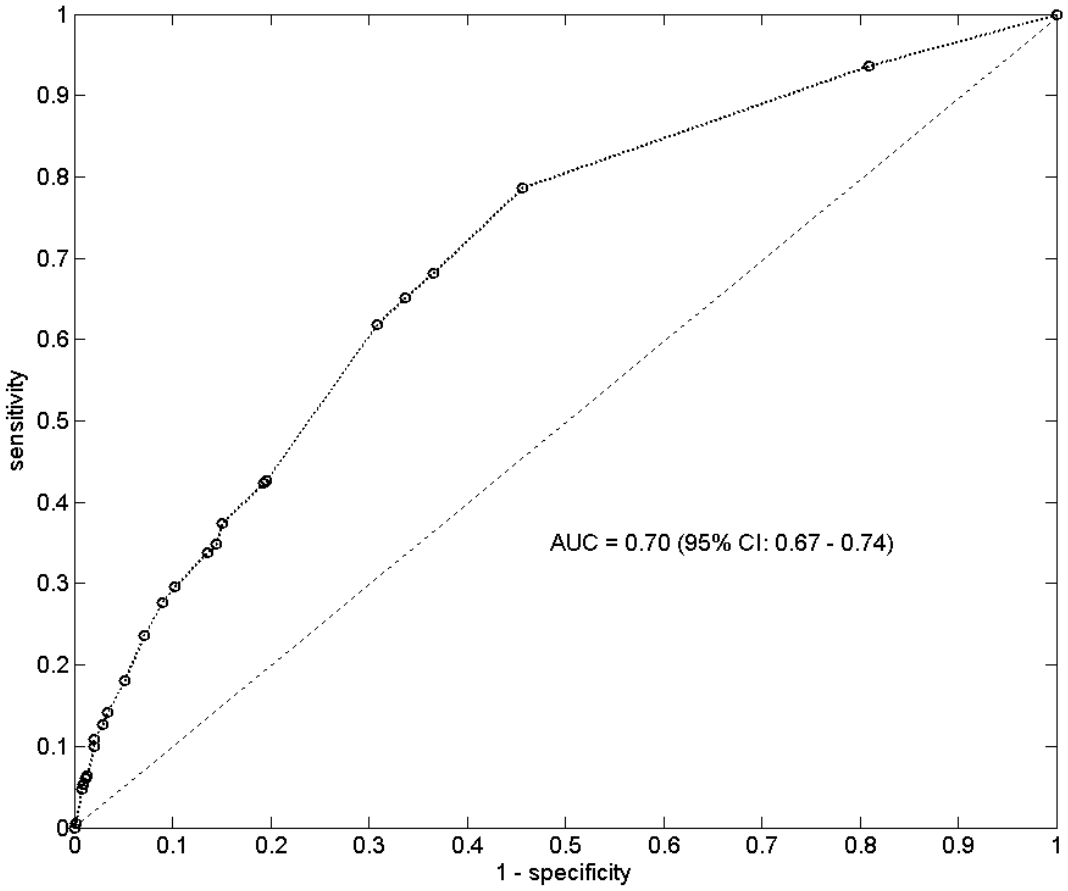

Figure 10 ROC curve of ER+/PR+ breast cancer prediction on test set

Figure 10 shows the ROC curve of the prediction results on the test set. The area under the ROC curve is 0.70, with 95% CI of 0.67 – 0.74. The actual data of Figure 10 can be found in Table 8. (In the ROC curve, sensitivity is defined as the number of breast cancer cases predicted as breast cancer divided by the total number of breast cancer cases, and specificity is defined as the number of non-breast cancer cases predicted as non-breast cancer divided by the total number of non-breast cancer cases. The model gives a probability between 0 and 1 for each record, and by varying the prediction threshold, we get a series of different sensitivity and specificity values, and thus the ROC curve. In application of the prediction model, one needs to pick a threshold, which is a trade-off between sensitivity and specificity.)

Both predicted and observed risks in Table 8 are prorated as discussed previously, to compensate for the sampling. We over-sampled breast cancer records, because for non-breast cancer nurses, we randomly selected one record; while for breast cancer nurses, we selected the record precedes the diagnosis of breast cancer. The proration ratio for all breast cancer is determined by dividing the expected number of breast cancer records if we had selected all records randomly by the actual number of breast cancer records. That is equal to dividing the actual number of breast cancer records (3216) by the total number of records of nurses with breast cancer (19186), which is 0.168. Then we also compensated for breast cancer cases missing ER/PR status, assuming in these cases the marginal distribution for ER+/PR+ breast cancer is the same as in those with known ER/PR status. The final proration ratio for ER+/PR+ breast cancer is 0.226, and we multiplied this ratio into both predicted risks and observed risks for a direct view of real-world risk.

Table 8 Prediction results of ER+/PR+ breast cancer

predicted risks (prorated)	observed risks (prorated)	number of records	observed cases	predicted cases		threshold	1- specificity	sensitivity
0.001	0.002	3451	23	17.255		0.001	1.000	1.000
0.003	0.002	6379	54	76.548		0.003	0.809	0.936
0.004	0.005	1675	38	28.475		0.004	0.457	0.787
0.005	0.005	514	11	10.280		0.005	0.365	0.681
0.006	0.005	530	12	13.250		0.006	0.337	0.651
0.006	0.007	2095	69	58.660		0.006	0.309	0.618
0.007	0.007	45	1	1.350		0.007	0.196	0.427
0.007	0.005	786	18	23.580		0.007	0.193	0.424
0.008	0.019	111	9	3.774		0.008	0.151	0.374
0.008	0.006	169	4	5.915		0.008	0.145	0.349
0.009	0.006	620	15	25.420		0.009	0.136	0.338
0.009	0.008	217	7	9.114		0.009	0.102	0.296
0.010	0.010	360	15	16.560		0.010	0.090	0.277
0.010	0.012	372	20	17.112		0.010	0.071	0.235
0.011	0.010	320	14	15.040		0.011	0.051	0.180
0.011	0.011	9	0	0.432		0.011	0.034	0.141
0.012	0.014	87	5	4.437		0.012	0.034	0.141
0.014	0.011	155	7	9.300		0.014	0.029	0.127
0.014	0.033	23	3	1.472		0.014	0.021	0.108
0.015	0.022	140	13	9.100		0.015	0.020	0.100
0.015	0.014	23	1	1.541		0.015	0.013	0.064
0.017	0.016	49	3	3.773		0.017	0.012	0.061
0.019	0.019	28	2	2.408		0.019	0.009	0.053
0.024	0.028	126	15	13.104		0.024	0.008	0.047
0.026	0.024	23	2	2.622		0.026	0.001	0.006
0.029	0.019	5	0	0.640		0.029	0.000	0.000

We also evaluated the model's ability to give an accurate probability for breast cancer prediction by comparing expected and observed incidence rates. The Bayesian Network predictor gives a probability to each record, as a predicted risk for that nurse. For nurses with the same input configuration, or variable values, the probability assigned by the Bayesian Network predictor will be the same. Therefore, the prediction results naturally group into clusters, with the same

predicted risk or probability for all records in the same cluster. For each cluster, we calculated the observed risk using a Bayesian approach based on the true class values of the records within that cluster, with uniform prior distribution and an equivalent sample size of 1. The size of the clusters (number of records within the cluster) varies over a wide range, as shown in Table 8, because the distribution of input configurations is unbalanced.

A comparison between predicted risk and observed risk is shown in Figure 11. The actual data are also in Table 8. The diagonal dotted line is an "ideal calibration curve," on which the results from a perfectly calibrated classifier would fall.

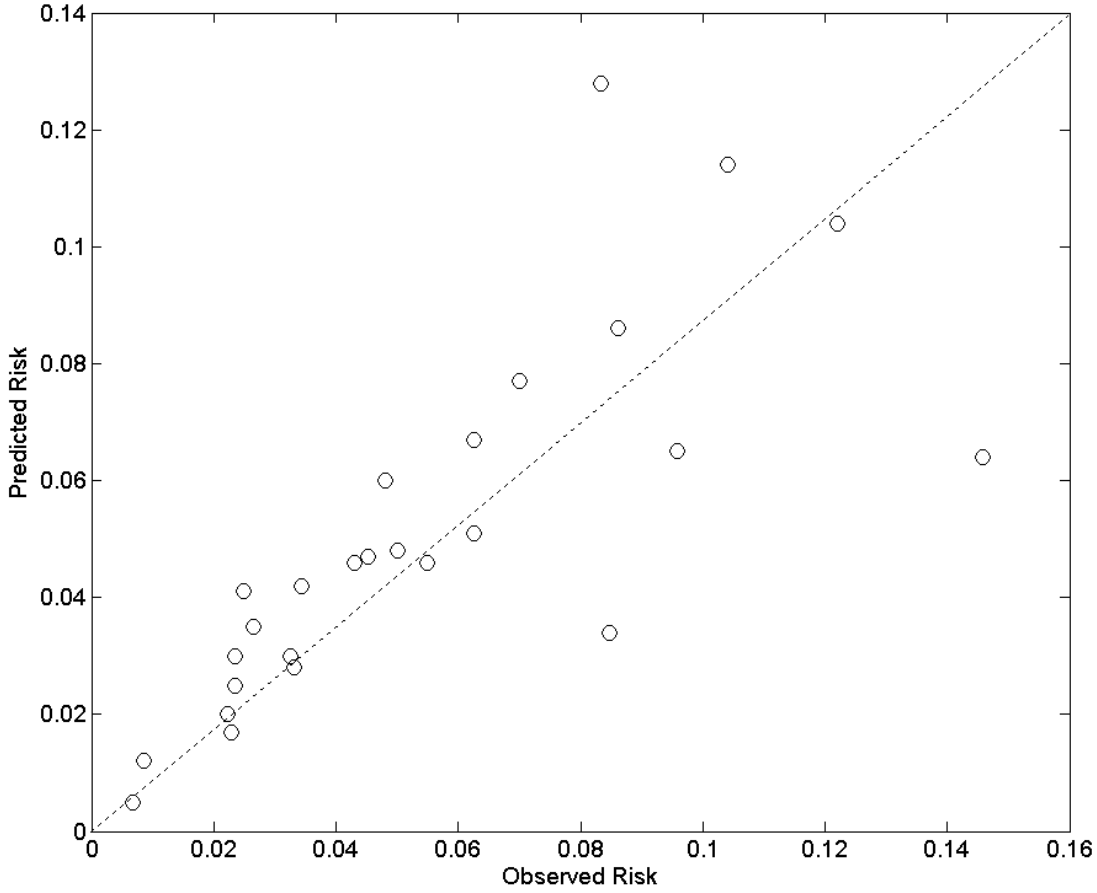

Figure 11 Calibration curve of ER+/PR+ breast cancer prediction on test set

There are two aspects we need to evaluate in this calibration curve. First, we want to know

whether there is a systematic overestimation or underestimation of risks. Second, we want to know, quantitatively, how close the predicted risks are to the observed risks on average. The first one can be evaluated by comparing the total expected number of breast cancer cases based on the predictions with the total observed number of breast cancer cases. The second one can be evaluated by a goodness-of-fit test.

For the above prediction results, the total expected number (E) of breast cancers is 370, which is the sum of expected breast cancer cases of every cluster, and the observed total number (O) is 361, giving an E/O ratio of 1.03. Rockhill *et al.* and Costantino *et al.* calculated CI of E/O by assuming a Poisson distribution for O. [38, 59] With this method, calculating the 95% CI of O and dividing E by these upper and lower numbers, the CI for above E/O is 0.93 – 1.14.

In a goodness-of-fit test, we calculated the χ^2 statistic of the predicted number of cases and observed number of cases by summing over the clusters, which gives χ^2 of 34.28 with corresponding p-value of 0.13, showing no statistically significant lack of fit.

Alternatively, the calibration curve can be evaluated using linear regression. The linear regression coefficient evaluates the closeness of predicted risks and observed risks, or degree of linear relationship between them. A linear regression on predicted risk versus observed risks gives a linear regression coefficient r^2 of 0.77. We aimed at improving this linear relationship between predicted and observed risks in next section.

4.6 Filtering of Risk Prediction Probabilities Based on Clustering

A closer look at the calibration curve shows that the data points that deviate further from the "ideal calibration curve" in Figure 11 generally have fewer records than the data points closer to the center. This observation prompted an attempt to improve the linearity of the calibration curve by clustering these data points.

Our Bayesian Network risk model gives the same probability prediction on records with the same input configuration, which constitute generic prediction clusters. The number of records within each generic cluster varies, reflecting the distribution of records over the input configurations. From the previous results, we see that clusters with more records are generally closer to the "ideal calibration curve" than clusters with fewer records. This observation can be intuitively understood, because with fewer records, the estimates of probabilities have larger variance.

Based on the above observation, we designed a filtering process to improve the linearity of the calibration curve. To be consistent with the model evaluation and validation process, this filter is also trained on the same training set on which we trained the model, and then applied to the prediction result of the same test set on which we tested the model. More specifically, we designed a filtering process that can learn a merging scheme from the training set predictions, i.e. which small clusters shall be merged together, and then apply this learned merging scheme to the test set prediction result, which looks like that we are filtering the prediction results such that the output only contains relatively larger or merged clusters.

If we merge the small clusters together, the resulting new calibration will deviate less from the "ideal" linear shape, but its discrimination ability may degrade, because the previously different or discriminating outputs now become the same. Therefore, we need to find a trade-off point that can maximize the linearity of the calibration curve without losing much discrimination power.

Slonim *et al.* developed a clustering algorithm that "explicitly maximizes the mutual information per cluster between the data and given categories," called the agglomerative information bottleneck method. [122] We employed a clustering approach based on this agglomerative information bottleneck method to find this trade-off point, as described below.

Suppose there are M_0 different input configurations in the data set, represented by x_{0i}, i = 1, 2 … M_0. Records with the same input configuration constitute a generic cluster, because merging these records (if we start from one cluster per record) would be trivial. The model gives probability prediction $p(y|x_{0i})$ on records within the same cluster x_{0i}, and $p(y|x_{0i})$ is the output for this cluster x_{0i}.

All generic clusters constitute the original cluster pool C_0. When two clusters x_a and x_b are merged into a new cluster x_z, x_a and x_b will be replaced by x_z in the cluster pool, and the cluster pool becomes C_1, which has $M_1 = M_0 - 1$ clusters. The output of the new cluster x_z is the weighted average of the output of x_a and x_b, i.e. $p(y \mid x_z) = \frac{p(x_a)p(y \mid x_a) + p(x_b)p(y \mid x_b)}{p(x_a) + p(x_b)}$. Then the merging procedure continues among the clusters within the cluster pool, while the size of the pool (number of current clusters) decreases.

The clustering is a greedy approach, merging two clusters together at a time. In each step, the clusters to be merged were selected from all possible combinations of any two clusters within the current cluster pool by the means of evaluating merge cost, which is weighted Jensen-Shannon divergence change.

Jensen-Shannon divergence (JS divergence) is a measurement of "distance" between conditional distributions. The JS divergence between two clusters can be calculated as following:

$$JS_{p(x1),p(x2)}[p(y \mid x1), p(y \mid x2)] = H\left[\sum_{i=1}^{2} p(x_i)p(y \mid x_i)\right] - \sum_{i=1}^{2} p(x_i)H[p(y \mid x_i)]$$

where *p(x1)* and *p(x2)* are the marginal distributions of two partitions, and *p(y|x1)* and *p(y|x2)* are the conditional probabilities of *y* given *x1* and *x2*, respectively. *H[p(x)]* is Shannon's entropy, given by $H[p(y)] = -\sum_y p(y) \log p(y)$.

Slonim *et al.* showed that in each merge step, the decrease in the mutual information between the clusters and the category variable (class variable) due to the merge is the JS divergence between the merged clusters weighted by the marginal distributions of these clusters, i.e.

$$merge\ cost = [p(x1) + p(x2)] \bullet JS_{p(x1),p(x2)}[p(y \mid x1), p(y \mid x2)]$$

By minimizing this "merge cost," we can find the "best possible merge" in each greedy step.

We applied this clustering procedure to the prediction results of the training set, and recorded during this procedure the area under the ROC curve and the linear regression coefficient of the observed and predicted risk, shown in Figure 12. From these results, we found a good trade-off point, when the number of clusters is 7, where both AUC and r^2 are very close to optimal values. The original AUC is 0.700, and the optimal r^2 is 1.000 with 3 clusters. With 7 clusters, the AUC is 0.698, and r^2 is 0.999. Please note these are prediction results on the training set, not the test set.

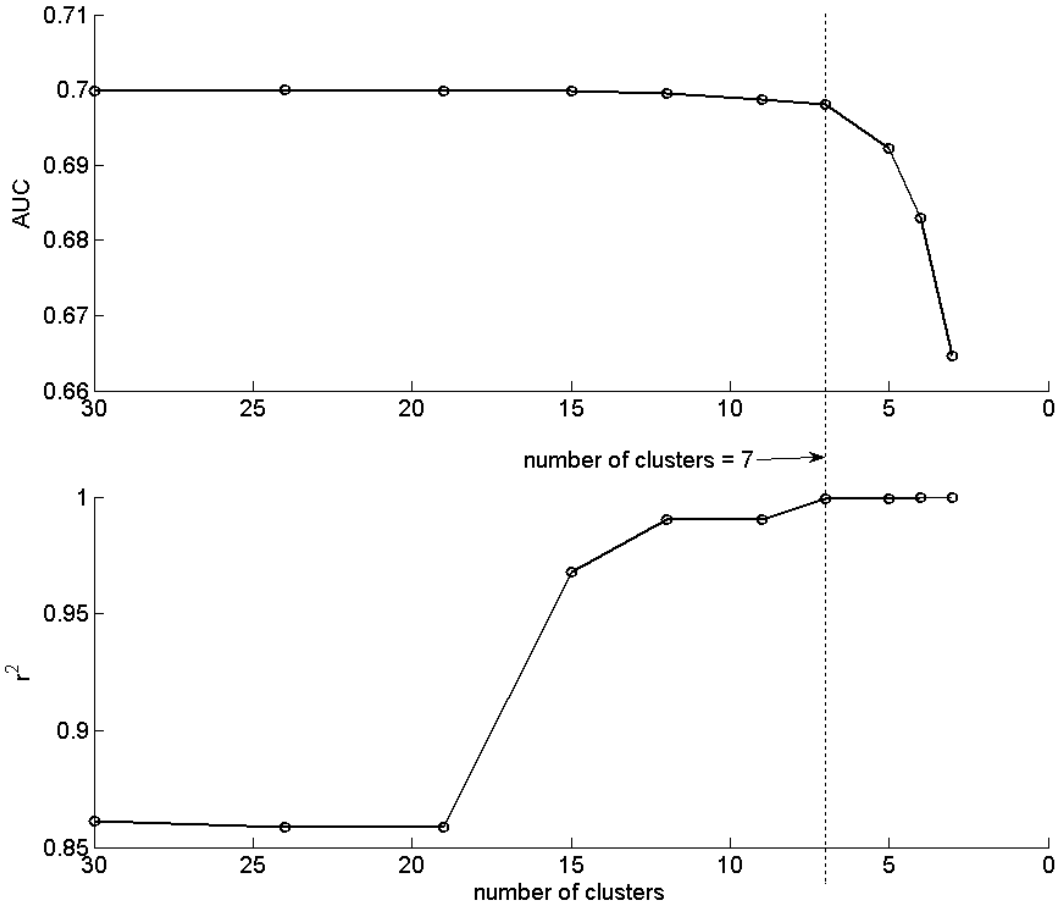

Figure 12 Clustering of prediction on training set

Therefore, we obtained a clustering filter, which can improve the calibration curve performance without losing much discrimination power in the training set, by mapping the model prediction output of each generic cluster to the prediction output of one of the seven clusters.

We then used this clustering filter, with number of clusters equal to 7, to smooth the prediction result of the test set. The prediction results after filtering are shown in Table 9

Table 9 Prediction results on test set after filtering

predicted risks (prorated)	observed risks (prorated)	count	observed cases	predicted cases		threshold	1-specificity	sensitivity
0.001	0.002	3451	23	17.255		0.001	1.000	1.000
0.003	0.002	6379	54	76.548		0.003	0.809	0.936
0.004	0.005	2189	49	38.865		0.004	0.457	0.787
0.006	0.007	3456	100	96.801		0.006	0.337	0.651
0.010	0.009	2265	89	97.129		0.010	0.151	0.374
0.015	0.016	418	29	27.778		0.015	0.029	0.127
0.024	0.026	154	17	16.304		0.024	0.008	0.047
						1	0	0

These prediction results after filtering give the ROC curve shown in Figure 13.

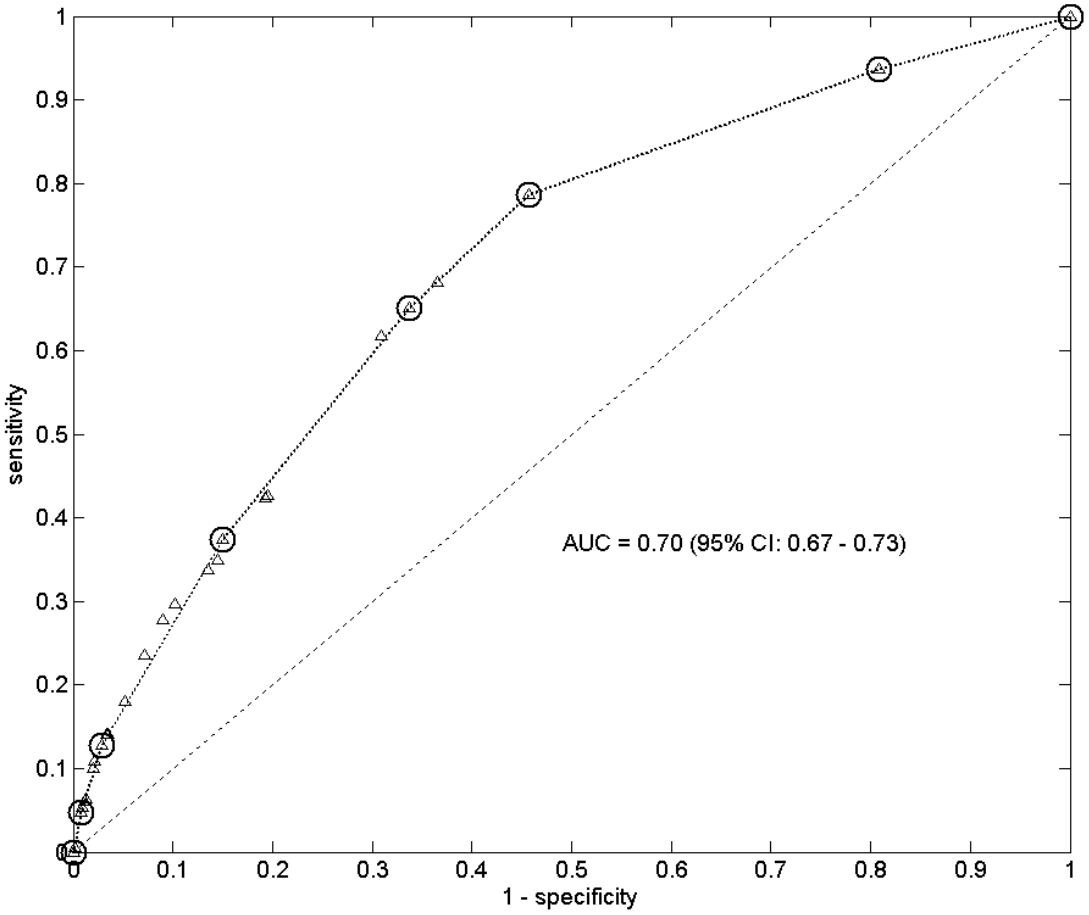

Figure 13 ROC curve after filtering on test set

In Figure 13, the circles connected by the dotted lines are data points after filtering, and the triangles are data points before filtering, i.e. the original clusters. From the graph, we see the clustering happened mostly among the clusters that are close neighbors (with similar sensitivity and specificity). The "crowded" areas at the low sensitivity end are now represented by one or two points, while the points without many neighbors at the high sensitivity end are kept. The AUC of this filtered result is 0.70, with 95% CI 0.67-0.73. Therefore, the clustering and filtering didn't lose much, if any, discrimination power compared to the original model.

The calibration curve of the prediction after filtering is shown in Figure 14 (the dotted line as the "ideal calibration curve" for comparison). From the graph, we can see now that the predicted risk

and observed risk have a higher degree of linear relationship than before clustering and filtering (in Figure 11). The linear regression coefficient r^2 is 0.996, while before clustering it was 0.77. We can see that in the test set the clustering filter also did a good job by improving linearity of the calibration curve and keeping the same discrimination power.

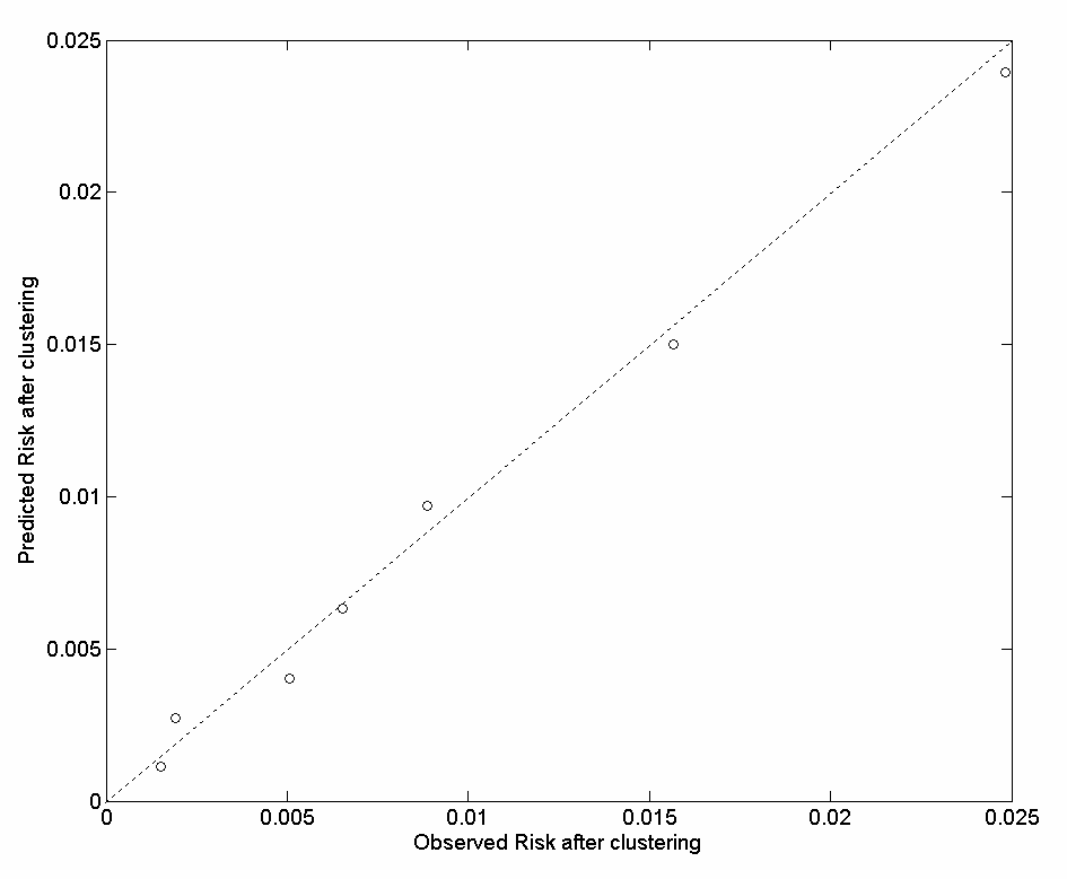

Figure 14 Calibration curve after filtering on test set

4.7 Classifiers Predicting Long Term Risks

The model discussed in the previous section, based on two-year interval data collection, predicts breast cancer incidence risk within two years. Clinically, longer term risk prediction may be desired for earlier and more effective prevention and disease management.

In order to build and test a model predicting breast cancer risks in the next N years, we need records of those nurses who were known not to develop breast cancer in the next N years as non-breast cancer records, and records of those nurses who were diagnosed of breast cancer randomly within the next N years as breast cancer records.

The non-breast cancer records can be obtained by randomly sampling one record from those at least N years before the last record, in which we know for sure a nurse didn't develop breast cancer within the next N years. The breast cancer records can be obtained by randomly sampling one record from within N years of the report of breast cancer diagnosis.

Putting these data together directly, however, may introduce a sampling bias that artificially increase prediction accuracy, because there will be no non-breast cancer records within the last N years of the study, but there could be breast cancer records during that period. Therefore, we need to filter the sampled data by discarding the breast cancer records that fall within the last N years of study. After such sampling and filtering, we get a 4-year risk data set with 1294 ER+/PR+ breast cancer nurses and 68308 non-breast cancer nurses, and a 6 year data set with 1176 ER+/PR+ breast cancer nurses and 64709 non-breast cancer nurses.

Again, by randomly splitting the data sets (stratified) into training sets with 75% data and test sets with the remaining 25% data, we learned Bayesian Networks from the training sets and evaluated the generalization errors on the test set. The Bayesian Networks for predicting 4-year and 6-year risks for ER+/PR+ breast cancer are shown in Figure 15 and Figure 16. In these two graphs, variable "Brcn" marks ER+/PR+ breast cancer versus non-breast cancer.

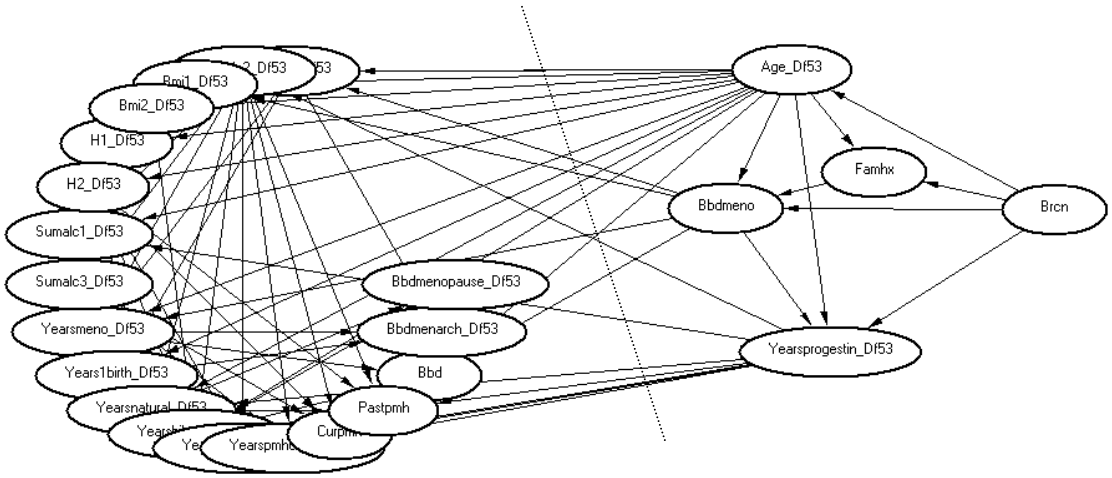

Figure 15 Bayesian Network for predicting 4-year risk of ER+/PR+ breast cancer

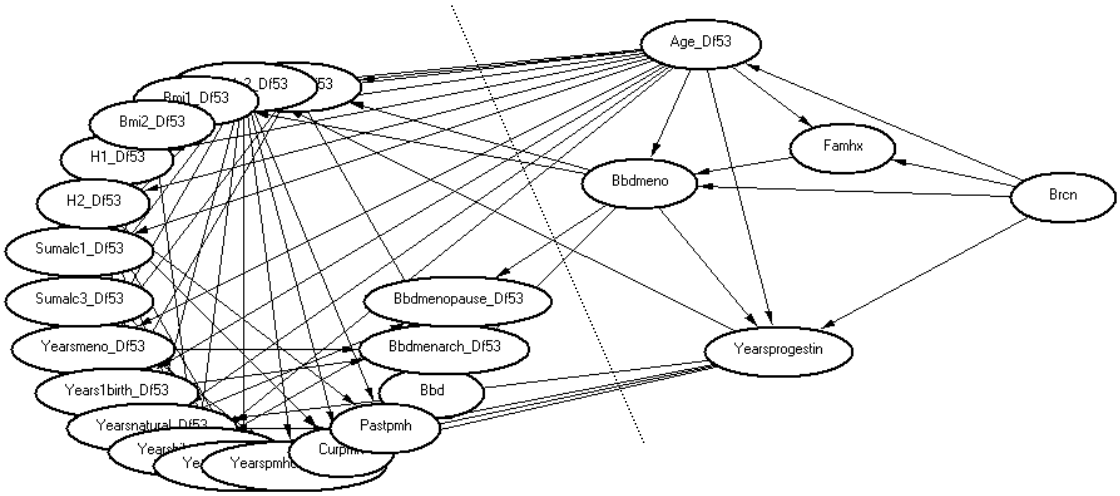

Figure 16 Bayesian Network for predicting 6-year risk of ER+/PR+ breast cancer

From the graphs, we can see the predicting variables are the same as in the 2-year risk model (Figure 8), age (Age_Df53), family history of breast cancer (Famhx), years having menstrual period if had benign breast disease (Bbdmeno), and years taking oral estrogen and progestin (Yearsprogestin). The prediction results are listed in Table 10.

Table 10 Prediction of 2-year, 4-year, and 6-year risk models for ER+/PR+ breast cancer

Model	AUC	95% CI of AUC	E/O ratio	95% CI of E/O ratio	p	r^2
2-year risk	0.70	0.67 – 0.74	1.03	0.93 – 1.14	0.13	0.77
4-year risk	0.68	0.64 – 0.72	1.02	0.91 – 1.14	0.75	0.79
6-year risk	0.66	0.62 – 0.69	1.06	0.94 – 1.19	0.55	0.28

We also applied filters constructed by agglomerative information bottleneck clustering on the predictions of 4-year and 6-year risks, and received similar improvement on the calibration curve without losing much discrimination power, shown in Table 11.

Table 11 Prediction after filtering of 2-year, 4-year, and 6-year ER+/PR+ risk models

Model	AUC	95% CI of AUC	E/O ratio	95% CI of E/O ratio	p	r^2
2-year risk	0.70	0.67 – 0.73	1.03	0.93 – 1.14	0.10	0.996
4-year risk	0.68	0.65 – 0.72	1.02	0.91 – 1.14	0.42	0.97
6-year risk	0.66	0.62 – 0.69	1.06	0.94 – 1.19	0.70	0.91

We developed models predicting 2-year, 4-year, and 6-year risks of ER+/PR+ breast cancer, and validated the models on hold-out test sets. The predicting structures are very similar for these three models, with age, family history of breast cancer, years on oral estrogen and progesterone use, and years having menstrual period given benign breast disease as the predicting factors for breast cancer.

We are not surprised to see such similarities between these models because of several reasons. First of all, the underlining biological and pathological mechanisms are the same, no matter whether we are predicting 2-year risks, 4-year risks, or 6-year risks. If a biological mechanism determines breast cancer risks in two years, it is possible that the same biological mechanism may determine breast cancer risks in four years as well, though we may expect to see more uncertainty

involved in longer term risks. Secondly, we sampled the breast cancer records in a way such that for 4-year risk data, about half of the breast cancer records are exactly the same as 2-year risk data, and for 6-year risk data, one third of the breast cancer records are exactly the same as 2-year risk data, and another one sixth are exactly the same as 4-year risk data (which is different from the 2-year data). In addition, because we discretized the variables, many of the different records actually got the same values after discretization except for the portion on the boundaries. These two effects together helped in deriving similar model structures. Thirdly, we learned these Bayesian Networks using the same variable orders, which favor similar structures.

Performance wise, our models have discrimination results in 2-year risk prediction better than those in 4-years, which in turn are better than those in 6-years. This is quite intuitive: longer term results generally involve more uncertainty, and thus are harder to predict. On the other hand, the differences between these models are not very large. Instead, there are large overlaps on the 95% CI of the AUCs between the results, and the AUC of 6-year risk prediction barely falls out of the 95% CI of the AUC of 2-year risk prediction. One reason for such small differences can be accredited to the similar predicting structures, while unchanged values caused by discretization might be another reason that can partially explain it.

4.8 Summary

In this chapter, we presented a series of models that we built to predict absolute risks of breast cancer over a defined period of time based on a set of variables constructed and selected by the log-incidence model proposed by Colditz *et al.*. [36] The data were randomly sampled for risk prediction of different period of time.

We evaluated the models on hold-out test sets, and reported the results in the form of AUC, calibration curve, E/O ratio, goodness-of-fit test, and linear regression of predicted and observed

risks. The AUC of the models is within the range from 0.66 to 0.70, and thus provides a relatively good discrimination power for breast cancer, given the fact that the current model used clinically, the Gail model, has an AUC of 0.58. [59] While a goodness-of-fit test showed no statistically significant lack of fit, the calibration curve actually showed rather good linear fit of the observed and predicted risks, and the overall E/O ratio is from 1.02 to 1.06.

In order to further improve the prediction performance on calibration, we constructed output filters by clustering the predictions on training sets, and then filtered the predictions on test sets using these filters. We reported the clustering results using the Agglomerative Information Bottleneck (AIB) method, which improved the linear relationship between the predicted and observed risks without sacrificing much discrimination power.

The discrimination performance of the models, measured by AUC, is far from a perfect model, which should have an AUC close to 1. Breast cancer is a disease that involves uncertainties. It is possible that whether a woman will develop breast cancer is uncertain until a very short time before the onset of the disease. In other words, clinical and life-style changes, spontaneous mutations in breast tissues, or other random factors might change the probability of developing breast cancer. It is possible that breast cancer is not a deterministic disease in the long term (e.g. two years in advance).

Because the biological and pathological mechanisms for breast cancer are unclear, or the exact information is unavailable, we used observable variables and risk factors based on past experience to build risk predicting models, which may be directly or indirectly linked to the true cause of breast cancer. For example, family history is linked to genetic defects and mutations, which should be one of the causes of breast cancer. Because we do not have the exact genetic information of all nurses, we used family history as a proxy to approximate the influence of the genetic factors. Such approximation brings in even more uncertainty in predicting breast cancer

compared to a hypothetical situation where we could directly use the genetic defects and mutations for prediction, which already involves uncertainty in low penetrance diseases such as breast cancer.

We are not saying that the prediction of breast cancer cannot be better. Some important risk factors, such as genotype of BRCA1 and BRCA2, were not available at the time of this study. Inclusion of such information will certainly boost the prediction power of the models. Also better models can be constructed when more advanced machine learning techniques are developed, or when we have a deeper understanding of the mechanism of breast cancer.

The calibration performance of the models, measured by E/O ratio, calibration curve, or linear relationship between observed and predicted risks, are quite good. Even with limited discrimination power, an accurate estimate of breast cancer risk is still useful, whether in clinical trial planning or disease counseling and management.

Chapter 5 Discussion

5.1 On Exploratory Analysis

During the procedure of exploratory analysis, we learned many lessons, which we wish we had known beforehand to work more efficiently and productively.

5.1.1 Expert Knowledge

Human expert knowledge is necessary in exploring complex dependencies even with advanced tools. One place we can use human expert knowledge is in variable selection. The search space of Bayesian Networks increases as a factorial of the number of variables, thus limiting variables in the study scope can help dramatically in limiting computation. In addition, incorporating expert knowledge in variable selection may help improving the performance of classification models. In our study, the classification models for ER+/PR+ breast cancer are learned from variables created from the original raw data based on human expert knowledge, i.e. the log-incidence model proposed by Colditz et al. with their understanding of the mechanism of breast cancer development. The classification performance of these models is better than that of other models we learned directly from the raw data. On the other hand, such selection of variables may impose unwarranted constraints on the parts of the hypothesis space being explored, especially when we try to find something of which nobody ever thought. There is always such a trade-off.

5.1.2 Exploration of the Dependencies in Learned Bayesian Networks

Learning a Bayesian Network from data sometimes prompts us with evidence of dependencies that we never thought of or have little knowledge of. One example is the dependency between clomid use and breast cancer with specific ER status discussed previously. This is one of the objectives of exploratory analysis: to find out things we don't know.

Domain specific knowledge and common sense are heavily involved in such explorations. Also the dependencies need to be explained with care. For instance, we found a dependency between breast implants and breast cancer in the process of exploration, such that nurses who had breast implants also had a higher incidence of breast cancer. After discussions with Professor Marco Ramoni and Professor Graham Colditz, we suspected that the dependency is probably due to breast implant after surgical removal of breast(s) because of breast cancer. To confirm this hypothesis, we further investigated this dependency by comparing the date of breast implant and the date of diagnosis of breast cancer. If the implant is after diagnosis of breast cancer, that record is re-marked as not having a breast implant. By such manipulation, we were able to evaluate breast implant as a possible risk factor for breast cancer. The result showed, on the contrary, that breast implant has a statistically insignificant association with lower incidence of breast cancer.

Such further exploration, guided by common sense and domain expert knowledge, is used to validate and confirm the dependencies found by learning a Bayesian Network from the data. To some extent, it can also help to understand the causal relationship that could be underlying the dependencies.

5.1.3 Split of Training and Test Set

The data sets we worked on are highly unbalanced, not only on breast cancer, but also on many other variables such as family history or clomid use. When splitting the data into a training set and a test set, we need to stratify the random splitting. For a more balanced data set, randomly splitting without stratification may not be a problem, because the random variance on the distribution is small comparing to a relatively large marginal distribution of the variable. For an unbalanced data set, the random noise can make a big difference. In our early analysis, we used non-stratified random splitting when training classification models. The prediction result has a heavy systematic bias by underestimating the risks. It turned out that was due to a higher incidence of breast cancer in the test set than the training set, which means we were evaluating a

model on a different population. In later analysis, we stratified all the random splitting and saw no more large systematic bias.

5.1.4 Discretization and Consolidation of Variable Values

The major tool we used, Bayesware Discoverer, requires continuous variables to be transformed into discrete variables. Many other Bayesian Network tools have the same requirement. Tools that can deal with continuous variables generally require the distribution of the continuous variable to be normal. Therefore, discretization is often a problem that we need to consider in Bayesian analysis.

There are many different ways to discretize continuous variables. Some machine learning algorithms can discretize during the learning process, such as C4.5, while many require it to be done in pre-processing the data. A continuous variable can be discretized based only on the information itself or it can be based on information from other variables. Commonly used methods to discretize a variable include binning to achieve equal numbers of data values in each bin, choosing bins that each span the same fraction of a variable's range, or bins based on normalized distributions. Other can be based on entropy or some statistics such as χ^2 values. There is no established theory about the best discretization method. In this work, we tried equal interval binning, equal size binning, and binning based on entropy and/or Jensen-Shannon divergence. We also tried to define the size of the bins to emphasize the extreme ends of values (e.g. top 1/5 and bottom 1/5 values as one bin each, while the middle 3/5 values as one bin). There is no statistically significant difference between these methods when comparing the prediction performance of the classification models we built.

5.2 On Learning Bayesian Networks from Data

5.2.1 Reading a Bayesian Network Learned From Data

Interpreting Bayesian Networks (BN) learned from data is different from using BN to represent human expert knowledge and reasoning. Learned BNs represents conditional independency structure. Knowledge based BNs generally represents causal relationships. Explaining the dependencies in a BN learned from data as causal relationship can be misleading or even dangerous.

The clomid analysis and breast implant analysis are two examples. In the clomid case, we checked for temporal succession to make sure that clomid was taken before the diagnosis of breast cancer, and matched clomid use on age and parity, two possible confounders. Therefore, with the resulting statistical association we suggest clomid as a possible risk factor for breast cancer with specific estrogen receptor status, or specifically, that it might increase the risk of ER+ breast cancer and decrease the risk of ER- breast cancer.

In the breast implant case, we first checked for temporal succession by comparing the date of breast implant with the date of diagnosis of breast cancer. It turned out that a large number of breast implants happened after breast cancer diagnosis. Then we marked these nurses as not having breast implant, so that the temporal relationship is consistent with a causal hypothesis as for a risk factor. The statistics after this manipulation showed statistically insignificant association between breast cancer and breast implant. All these analyses together illustrate that the dependency we found in the original Bayesian Network reveals a possible causal relationship, but with breast cancer as the cause and breast implant as the result.

We also need always to keep in mind that the dependency structure we see is only one of a group of equally or possibly even more probable structures. We cannot make any conclusion without a reasonable amount of effort, and even then the conclusion is only suggestive, not definitive, as we've seen in the d-separation analysis of the risk score model.

5.2.2 Bayesian Network and Highly Interdependent Data

With an infinite amount of data, we can build a joint probability distribution table as a predictor with the best possible performance, i.e. no better prediction can be made without extra variable information. In reality, however, it is always the case that our data are very limited. Even with tens of thousands of records, we still soon run out of data as the number of data points necessary to make a reasonably good estimation increases exponentially with the number of variables (and their values) in the joint probability distribution table.

People have been using Bayesian Networks to deal with such problems by reducing the parameter space with a well defined dependency structure, and in many cases have been successful. However, if the variables in the data set are highly interdependent, we will still encounter difficulties.

For example, for a data set with n binary variables, 2^n parameters are necessary to fully describe the joint probability distribution table. If the underlying dependency structure can be represented with a Bayesian Network in which all nodes are isolated, i.e. all independent of each other, we need only n parameters to fully describe the distribution. That's a reduction of parameters from exponential to linear. If the underlining dependency structure can be represented with a Bayesian Network in which every node has exactly one parent and one child, except for the top one and the bottom one, we need 2n-1 parameters. More generally, a Bayesian Network with n nodes, each has p_i number of parents, will need $\sum_{i=1}^{n} 2^{p_i}$ number of parameters to fully describe, while p_i is constrained by the acyclic restriction. When the Bayesian Network is a complete graph, this number becomes 2^n, which is the same as that of a joint probability distribution table.

If the underlying dependency structure is actually complex, the number of dependencies needed

in a BN to describe it grows, and the amount of data needed to learn these additional parameters also grows. When the data are insufficient, we will end up with a simpler structure, reflecting inadequate information to override the independence assumptions.

Therefore, learning a good Bayesian Network from data becomes harder when the interdependency among the variables increases. In extreme cases, it can be as hard as estimating a joint probability distribution table.

In such situations, we will have to seek methods to reduce the interdependencies among the data. Merging variables and consolidating variable values could be useful, yet our limited effort on this problem didn't make much difference in the experiments reported here. This forms one interesting problem for future discussion.

5.3 On Evaluation of Risk Predicting Models

5.3.1 Evaluation of Models

We need to clarify two factors in evaluating a risk prediction model. One is the prediction capability of the model itself, or how well the model captures the structure of the data without overfitting. This is generally performed by evaluating the generalization error of the model. The other is the representativeness of the population on which the model is built. If the population used to build the model is not representative, the generalization ability of the model will be limited.

Evaluating the model using hold-out test samples, cross-validation, calibration curves, or other similar methods gives a measure of the prediction capability of the model itself. Comparing the marginal distributions of various variables with those in a "gold standard" population, e.g. the whole population of the US, can give a rough idea on the representativeness of the population.

Evaluating the model on a different population, for instance, a different study, can be a combination of both, but such efforts need to be interpreted with care because this other test population may not be representative as well.

In this work, we evaluated the prediction capability of the model using a hold-out test set, and therefore applying these models to any population other than NHS needs careful checking on the target population distribution and probably even model adjustments.

We evaluated the models on discrimination using AUC, and on calibration using a calibration curve, E/O ratio, goodness-of-fit test, and linear regression. The AUC of the models is within the range from 0.66 to 0.70, and thus provides a relatively good discrimination power for breast cancer. Considering calibration, all models passed a goodness-of-fit test with no statistically significant lack-of-fit, and with a good linear relationship between predicted and observed risks, especially after clustering.

Even though the discrimination power is limited (best AUC of 0.70, good for breast cancer, but still limited), a well-estimated risk can still be useful in certain applications such as deciding whether to administer a preventive intervention that has adverse and beneficial effects.

5.3.2 Comparison with the Model in Clinical Use

The Gail model is a clinically applied breast cancer risk model, and evaluation results of this model are available. Spiegelman *et al.* reported a linear regression coefficient of 0.67 between observed and predicted risks. [58] Costantino *et al.* reported overall E/O ratio 0.84 (95% CI 0.73-0.97) on Gail model 1 and overall E/O ratio 1.03 (95% CI 0.88-1.21) on Gail model 2. [38] Rockhill *et al.* reported AUC of 0.58 (95% CI 0.56-0.60), and overall E/O ratio 0.94 (95% CI 0.89-0.99) on Gail model 2. [59]

Comparing models presented in this work with the Gail model, however, is infeasible at the current stage of our work. This is because of two reasons. First, the Gail model predicts the risk for all breast cancer, while the majority efforts of our work focused on ER+/PR+ breast cancer, because it is clinically important and useful. Second, the Gail model predicts breast cancer risk in a five year period (in clinical applications as well as the published evaluation results), while the models of this work predict risks in two, four, or six year periods. Measures can be taken to make these models comparable to the Gail model, which will be discussed later.

5.3.3 Clustering of Bayesian Network Prediction

The data set we used is highly unbalanced, not only in the class variable, breast cancer, but also in predicting variables such as family history and hormone use. It is common that some of the prediction buckets are much smaller than others. Predictions on these small buckets are generally inaccurate. Clustering the small buckets together may help improve the robustness of the prediction. The reason for improvement in calibration but not in AUC is due to the small size of the inaccurate buckets, which are clustered in the process. For overall accuracy, such changes may not have much impact. However, when applied in clinical decision making, for the individuals falling into those small clusters, having a better estimation of risk (probability) can make a significant impact on the final medical decision.

We reported the clustering results using the Agglomerative Information Bottleneck (AIB) method, which improved the linear relationship between the predicted and observed risks without sacrificing much discrimination power. During the exploratory analysis, we also tried clustering based on the Euclidean distance between the inputs, with or without penalizing on the Jensen-Shannon distance, and ad hoc clustering simply based on the output. We finally decided to use AIB because it has a well-established theory and better empirical performance, i.e. faster increase in linear relationship and slower decrease on AUC.

5.4 Summary of Contributions

In this thesis, we performed exploratory analysis on a rich collection of data from a large health cohort study, the Nurses' Health Study.

In Chapter 3, we illustrated with examples how Bayesian Networks can be used to explore the dependencies among a group of variables, and how to exercise care when explaining the found dependencies and validate for causal relationship. We investigated the dependency structures among a collection of clinical and life-style variables and a selected set of SNPs from different genes, and obtained a null result that these SNPs may not be associated with breast cancer. We also found during the exploration a new risk factor, clomid use, for breast cancer with specific estrogen receptor status.

In Chapter 4, we first attempted to evaluate risk score as an index for breast cancer incidence, and then we developed risk prediction models using Bayesian Networks, first on all breast cancer, then on ER+/PR+ breast cancer, which is more prone to the current preventive intervention and hence whose prediction is clinically more significant. We evaluated the performance of the risk prediction models on discrimination ability using AUC, which is good compared to the Gail model used clinically. We also evaluated the calibration of the models using calibration curves, $\chi 2$ statistics, E/O ratio, and linear relationship of predicted and observed risks. We applied Agglomerative Information Bottleneck clustering to construct a prediction filter for the risk model, and improved the calibration performance without sacrificing much discrimination power by applying the filter to prediction outputs.

Overall, our contributions include the discovery of a new risk factor for breast cancer with specific estrogen receptor status, and the development of prediction models for breast cancer incidence risks in different periods of time. We applied prediction output filters to the models we

developed, and showed improved calibration performance, which is clinically important in disease management and clinical trial planning. We also showed that with a reasonable amount of effort, no dependency could be found between a group of SNPs and breast cancer, suggesting no association exists between the related genes and breast cancer.

5.5 Future Work

Exploratory analysis on a large health cohort study such as the Nurses' Health Study can be unending. The work presented in this thesis is very limited, and has a large potential for further improvement.

5.5.1 Data Pre-processing

Data collected in the Nurses' Health Study don't have an overwhelming number of missing values, and we didn't suffer much from missing values in this work. We used k-nearest neighbor imputation to deal with the missing values. There are many superior methods to deal with missing values, and recruiting those methods may render better results.

We already discussed discretization and consolidation of variable values. We tried different ways of discretization, but there are more options available. For example, we can discretize continuous variables in the process of learning a Bayesian Network, which we didn't try due to the limitation of the tools we used. In the latest exploration, we tried discretization on Jensen-Shannon divergence, which showed some signs of improvement. This could be another promising direction for future work.

5.5.2 Evaluation of Risk Score As an Index for Breast Cancer

The work we presented on evaluating the risk score as an index for breast cancer is incomplete.

We believe that a similar analysis could be completed based on all the data that were used in developing the log-incidence model, which were not all available to us in the present study. With those additional data, we could search a wide range of possible Bayesian Network structures to find the best predictive models and to assess our confidence in them.

5.5.3 Dealing with Highly Interdependent Data

Learning Bayesian Networks is difficult with highly interdependent data. There could be other ways to help attack this problem. For example, we can try to use a model with latent variables, hoping the introduction of a latent variable can reduce the interdependencies. Alternatively, we can construct a group of Bayesian Networks from the data with different learning heuristics or parameters, and compose a Bayesian Committee with this group of Bayesian Networks. Such a meta-classifier may have more robust performance.

5.5.4 Prediction of 5-year Risks

Clinically, people are more used to estimating 5-year risks, partly due to the tradition. The models presented in this work predict two, four, or six year risks, because the data were collected in two year intervals. With time alignment and interpolation of the data, however, it is possible to develop 5-year risk models using the same approach as in this work. In order to make this work applicable in a clinical setting, probably developing a 5-year risk model is necessary.

The development of a 5-year risk model will also help to make it comparable with the Gail model, if clinical guidance can be established to project ER+/PR+ risks to all breast cancer risks.

5.5.5 Clomid and Breast Caner

Clomid, or clomiphene citrate, has been widely used in treatment for fertility enhancement since

the early 1970's, because of its ease of administration and minimal side effects. It acts as a selective estrogen receptor modulator, binding to estrogen receptors and has mixed agonist and antagonist activity depending upon the target tissue. [113] Clomid works as an estrogen antagonist at the hypothalamus, causing it to sense low levels of serum estrogen and increase the production of gonadotropin releasing hormone, which stimulates the pituitary gland to produce follicle stimulating hormone (FSH). FSH stimulates the development of the ovarian follicles, which contain the eggs. In addition, clomid works on the pituitary gland to boost the production of Luteinizing Hormone, which triggers the ovulation process and maturation of the eggs. [114]

Clomid works as an estrogen agonist on some tissues, while estrogen and estrogen-like hormones are known to be risk factors for breast cancer. Is this why clomid use is associated with higher risk of ER+ breast cancer? If so, are there similar effects of other estrogen-like hormones, i.e., do they specifically raise the risk of ER+ breast cancer?

Women take clomid as a fertility drug, which happens during reproductive age, or at least before menopause. Women take PMH for hormone replacement therapy, generally after menopause. While tamoxifen is generally used for prevention and treatment of early stage breast cancer, which is more likely to happen in the later part of a women's life. Could this different stage of life when taking the hormones have an impact on the outcome?

Clomid causes many different hormonal changes. To name a few, it increases androstenedione, dehydroepiandrosterone, dehydroepiandrosterone sulfate, dihydrotestosterone, estradiol, estrogens (urine), follicle stimulating hormone, luteinizing hormone, progesterone, testosterone, free testosterone, thyroxine binding globin, thyroid stimulating hormone, and decreases cholesterol, triiodothyronine and free thyroxine index. [115] Could any of these changes, or any other not listed here, be related to breast cancer risks?

All these questions cannot be answered without further investigation and form interesting topics for future work.

Bibliography

1. Nurses' Health Study. Available at: http://www.channing.harvard.edu/nhs/hist.html.
2. R. Chen, personal communication.
3. Chan AT, Giovannucci EL, Meyerhardt JA, Schernhammer ES, Curhan GC, Fuchs CS. Long-term Use of Aspirin and Nonsteroidal Anti-inflammatory Drugs and Risk of Colorectal Cancer. JAMA. 2005;294:914-923.
4. The Nurses' Health Study (NHS). Available at: http://www.channing.harvard.edu/nhs/index.html
5. Cotran RS, Kumar V, Collins T. Pathologic Basis of Disease, 6th Edition. W.B. Saunders Company, 1999.
6. American Cancer Society. Available at: http://www.cancer.org/docroot/STT/stt_0_2004. asp?sitearea=STT&level=1.
7. Bross DJ, Blumenson LE. Statistical testing of a deep mathematical model for human breast cancer. J Chronic Dis. 1968 Dec; 21(7):493-506.
8. Hilf R. Will the Best Model of Breast Cancer Please Come Forward. Natl Cancer Inst Monogr. 1971 Dec; 34:43-54.
9. MacMahon B, Cole P, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S. Age at first birth and breast cancer risk. Bull World Health Organ. 1970; 43(2):209-21.
10. Staszewski J. Age at menarche and breast cancer. J Natl Cancer Inst. 1971 Nov; 47(5):935-40.
11. Trichopoulos D, MacMahon B, Cole P. Menopause and breast cancer risk. J Natl Cancer Inst. 1972 Mar; 48(3):605-13.
12. MacMahon B, Morrison AS, Ackerman LV, Lattes R, Taylor HB, Yuasa S. Histologic characteristics of breast cancer in Boston and Tokyo. Int J Cancer. 1973 Mar 15; 11(2):338-44.
13. De Waard F, Cornelis JP, Aoki K, Yoshida M. Breast cancer incidence according to weight and height in two cities of the Netherlands and in Aichi prefecture, Japan. Cancer. 1977 Sep; 40(3):1269-75.
14. Farewell VT. The combined effect of breast cancer risk factors. Cancer. 1977 Aug; 40(2):931-6.
15. Anderson DE, Badzioch MD. Risk of familial breast cancer. Cancer. 1985 Jul 15; 56(2):383-7.
16. Henderson BE, Pike MC, Casagrande JT. Breast cancer and the oestrogen window hypothesis. Lancet. 1981 Aug 15;2(8242):363-4.
17. Thomas DB, Persing JP, Hutchinson WB. Exogenous estrogens and other risk factors for breast cancer in women with benign breast diseases. J Natl Cancer Inst. 1982 Nov;69(5):1017-25.
18. Thomas DB. Factors that promote the development of human breast cancer. Environ

Health Perspect. 1983 Apr; 50:209-18.
19. Lipsett MB. Hormones, medications, and cancer. Cancer. 1983 Jun 15; 51(12 Suppl):2426-9.
20. Trichopoulos D, Hsieh CC, MacMahon B, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S. Age at any birth and breast cancer risk. Int J Cancer. 1983 Jun 15;31(6):701-4.
21. Polednak AP, Janerich DT. Characteristics of first pregnancy in relation to early breast cancer. A case-control study. J Reprod Med. 1983 May ;28(5):314-8.
22. Moolgavkar SH, Day NE, Stevens RG. Two-stage model for carcinogenesis: Epidemiology of breast cancer in females. J Natl Cancer Inst. 1980 Sep;65(3):559-69.
23. Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG. Hormonal risk factors, breast tissue age and the age-incidence of breast cancer. Nature. 1983 Jun 30; 303(5920):767-70.
24. Ottman R, Pike MC, King MC, Henderson BE. Practical guide for estimating risk for familial breast cancer. Lancet. 1983 Sep 3; 2(8349):556-8.
25. Claus EB, Risch N, Thompson WD. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. Breast Cancer Res Treat 1993; 28:115-20.
26. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer 1994 ; 73:643 – 51.
27. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989 Dec 20;81(24):1879-86.
28. Anderson SJ, Ahnn S, Duff K. NSABP Breast Cancer Prevention Trial risk assessment program, version 2. NSABP Biostatistical Center Technical Report, August 14, 1992.
29. Couch FJ, DeShano ML, Blackwood MA, et al: BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. N Engl J Med 336:1409-1415, 1997.
30. Shattuck-Eidens D, Oliphant A, McClure M, et al: BRCA1 sequence analysis in women at high risk for susceptibility mutations: Risk factor analysis and implications for genetic testing. J Am Med Assoc 278:1242-1250, 1997.
31. Frank TS, Manley SA, Olopade OI, et al: Sequence analysis of BRCA1 and BRCA2: Correlation of mutations with family history and ovarian cancer risk. J Clin Oncol 16:2417-2425, 1998.
32. Pathak DR, Whittemore AS. Combined effects of body size, parity, and menstrual events on breast cancer incidence in seven countries. Am J Epidemiol. 1992 Jan 15;135(2):153-68.
33. Parmigiani G, Berry D, Aguilar O: Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. Am J Hum Genet 62:145-158, 1998.
34. Rosner B, Colditz GA. Nurses' health study: log-incidence mathematical model of breast cancer incidence. J Natl Cancer Inst 1996; 88: 359 – 64.
35. Colditz GA, Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. Am J Epidemiol 2000 ; 152 : 950 – 64.

36. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk factors for breast cancer according to estrogen and progesterone receptor status. J Natl Cancer Inst 2004; 96: 218 – 28.
37. Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG. Hormonal risk factors, breast tissue age and the age-incidence of breast cancer. Nature 1983; 303:767-70.
38. Costantino J, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999; 91:1541–8.
39. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004 ; 23 : 1111 – 30.
40. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66-71.
41. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789-92.
42. Tavtigian SV, Simard J, Rommens J, et al. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. Nat Genet 1996;12: 333-7.
43. Berry DA, Iversen ES Jr, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, et al. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. J Clin Oncol 2002 ; 20 : 2701 – 12.
44. Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, et al. Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10 000 individuals. J Clin Oncol 2002 ; 20 : 1480 – 90.
45. Antoniou AC, Pharoah PD, McMullan G, Day NE, Stratton MR, Peto J, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer 2002 ; 86 : 76 – 83.
46. de la Hoya M, Osorio A, Godino J, Sulleiro S, Tosar A, Perez-Segura P, et al. Association between BRCA1 and BRCA2 mutations and cancer phenotype in Spanish breast/ovarian cancer families: implications for genetic testing. Int J Cancer 2002 ; 97 : 466 – 71.
47. de la Hoya M, Diez O, Perez-Segura P, Godino J, Fernandez JM, Sanz J, Alonso C, Baiget M, Diaz-Rubio E, Caldes T. Pre-test prediction models of BRCA1or BRCA2mutation in breast/ovarian families attending familial cancer clinics. J. Med. Genet 2003; 40:503-10.
48. Vahteristo P, Eerola H, Tamminen A, Blomqvist C, Nevanlinna H. A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. Br J Cancer 2001 ; 84 : 704 – 8.
49. Hartge P, Struewing JP, Wacholder S, Brody LC, Tucker MA. The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. Am J Hum Genet 1999 ; 64 : 963 – 70.
50. Apicella C, Andrews L, Hodgson SV, Fisher SA, Lewis CM, Solomon E, et al. Log odds of carrying an ancestral mutation in BRCA1 or BRCA2 for a defi ned personal and family history in an Ashkenazi Jewish woman (LAMBDA). Breast Cancer Res 2003 ; 5 :

R206 – 16.
51. Jonker MA, Jacobi CE, Hoogendoorn WE, Nagelkerke NJ, de Bock GH, van Houwelingen JC. Modeling familial clustered breast cancer using published data. Cancer Epidemiol Biomarkers Prev 2003 ; 12 : 1479 – 85.
52. Gilpin CA, Carson N, Hunter AG. A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center. Clin Genet 2000 ; 58 : 299 – 308.
53. Fisher TJ, Kirk J, Hopper JL, Godding R, Burgemeister FC. A simple tool for identifying unaffected women at a moderately increased or potentially high risk of breast cancer based on their family history. Breast 2003;12:120 –7.
54. Wingo PA, Ory HW, Layde PM, Lee NC. The evaluation of the data collection process for a multicenter, population-based, case–control design. Am J Epidemiol 1988;128:206–17.
55. Gail MH, Benichou J. Assessing the risk of breast cancer in individuals. In: DeVita VT Jr, Hellman S, Rosenberg SA, editors. Cancer prevention. Philadelphia (PA): Lippincott; 1992. p. 1–15.
56. Gail MH, Benichou J. Validation studies on a model for breast cancer risk [editorial] [published erratum appears in J Natl Cancer Inst 1994;86:803]. J Natl Cancer Inst 1994;86:573–5.
57. Bondy ML, Lustbader ED, Halabi S, Ross E, Vogel VG. Validation of a breast cancer risk assessment model in women with a positive family history. J Natl Cancer Inst 1994;86:620–5.
58. Spiegelman D, Colditz GA, Hunter D, Hertzmark E. Validation of the Gail et al. model for predicting individual breast cancer risk. J Natl Cancer Inst 1994;86:600–7.
59. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001;93:358-66.
60. Yasui Y, Potter JD. The shape of the age-incidence curves of female breast cancer by hormone-receptor status. Cancer Causes Control 1999;10:431–7.
61. Tarone RE, Chu KC. The greater impact of menopause on ER- than ER+ breast cancer incidence: a possible explanation (United States). Cancer Causes Control 2002;13:7–14.
62. Hulka BS, Chambless LE, Wilkinson WE, Deubner DC, McCarty KS Sr, McCarty KS Jr. Hormonal and personal effects of estrogen receptors in breast cancer. Am J Epidemiol 1984;119:692–704.
63. Hildreth NG, Kelsey JL, Eisenfeld AJ, LiVolsi VA, Holford TR, Fischer DB. Differences in breast cancer risk factors according to the estrogen receptor level of the tumor. J Natl Cancer Inst 1983;70:1027–31.
64. Wolpert D. Stacked generalization. Neural Networks, 1992; 5 : 241 – 59.
65. Breiman L. Bagging predictors. Machine Learning, 1996; 24(2):123 - 40.
66. Breiman L. Random forest. Machine Learning, 2001; 45: 5 – 32.
67. Dempster AP, Laird D, Rubin D. Maximum likelihood from incomplete data via the EM

algorithm (with discussion). Journal of the Royal Statistical Society 1977; 39:1-38.
68. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE transactions on Pattern Analysis and Machine Intelligence 1984; 6:721-41.
69. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, 1988.
70. Chickering DM. Learning Bayesian networks is NP-Complete. Learning from Data: Artificial Intelligence and Statistics, pp. 121-130. Springer-Verlag, 1996.
71. Pacific Fertility Center.
Available at: http://www.infertilitydoctor.com/treat/clomiphene.htm
72. Jordan M. Learning in Graphical Models. MIT Press, Cambridge, MA, 1999.
73. Bayesware Limited.
Available at: http://www.bayesware.com/products/discoverer/discoverer.html
74. Cooper G, Herskovitz E. A bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992; 309–347
75. Ramoni M. Personal communication.
76. Dixon JK. Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics, 1979; SMC-9, 10, 617-621.
77. Duda, RO., Hart, PE., and Stork, DG.. Pattern Classification 2nd, John Wiley & Sons, 2001.
78. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Bostein D. and Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics, 2001; 17(6): 520-5.
79. Tago C, Hanai T. Prognosis prediction by microarray gene expression using support vector machine. Genome Informatics, 2003; 14: 314-5.
80. National Human Genome Research Institute. Introduction to the Human Genome Project, [online] available http://www.genome.gov/10001772, 2003.
81. Thompson MW, McInnes RR, and Willard HF. Thompson & Thompson: Genetics in Medicine, Fifth Edition, W. B. Saunders Company, Philadelphia, 1991.
82. Wang DG, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science, vol. 280, no. 5366, pp. 1077-1082, 1998.
83. Altshuler D, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature, vol. 407, no. 6803, pp. 513-516, 2000.
84. Mullikin JC, et al. An SNP map of human chromosome 22. Nature, vol. 407, no. 6803, pp. 516-520 2000.
85. Marth GT, et al. A general approach to single-nucleotide polymorphism discovery. Nature Genetics, vol. 23, no. 4, pp. 452-456, 1999.
86. Botstein D, and Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics Supplement, vol. 33, pp. 228-237, 2003.

87. Environmental Genome Project. Polymorphism and genes. [online] available http://www.niehs.nih.gov/envgenom/polymorf.htm, 2003.
88. Hall JM, et al. Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21. Science, vol. 250, no. 4988, pp. 1684-1689, 1990.
89. Miki Y, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science, vol. 266, no. 5182, pp. 66-71, 1994.
90. Wooster R, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature, vol. 278, no. 6559, pp. 789-792, 1995.
91. Serova OM, et al. Mutations in BRCA1 and BRCA2 in breast cancer families: are there more breast cancer-susceptibility genes? American Journal of Human Genetics, vol. 60, no. 3, pp. 486–495, 1997.
92. Seitz S. Strong indication for a breast cancer susceptibility gene on chromosome 8p12-p22: linkage analysis in German breast cancer families. Oncogene, vol. 14, no. 6, pp. 741-743, 1997.
93. de Jong MM, et al. Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. Journal of Medical Genetics, vol. 39, no. 4, pp. 25-242, 2002.
94. Helzlsouer KJ, et al. Association between glutathione S-transferase M1, P1, and T1 genetic polymorphisms and development of breast cancer. Journal of the National Cancer Institute, vol. 90, no. 7, pp. 512-518, 1998.
95. Charrier J**,** et al. Allelotype influence at glutathione S-transferase M1 locus on breast cancer susceptibility. British Journal of Cancer, vol. 79, no. 2, pp. 346-353, 1999.
96. Ambrosone CB, et al. Cytochrome P4501A1 and glutathione S-transferase (M1) genetic polymorphisms and postmenopausal breast cancer risk. Cancer Research, vol. 55, no. 16, pp. 3483-3485, 1995.
97. Kristensen VN, et al. A rare CYP19 (aromatase) variant may increase the risk of breast cancer. Pharmacogenetics, vol. 8, no. 1, pp. 43-48, 1998.
98. Haiman CA, et al. No association between a single nucleotide polymorphism in CYP19 and breast cancer risk. Cancer Epidemiology Biomarkers & Prevention, Vol. 11, no. 2, pp. 215-216, 2002.
99. Haiman CA, et al. A tetranucleotide repeat polymorphism in CYP19 and breast cancer risk. International Journal of Cancer, vol. 87, no. 2, pp. 204-210, 2000.
100. Bailey LR, et al. Breast cancer and CYPIA1, GSTM1, and GSTT1 polymorphisms: evidence of a lack of association in Caucasians and African Americans. Cancer Research, vol. 58, no. 1, pp. 65-70, 1998.
101. Huang CS, et al. Breast cancer risk associated with genotype polymorphism of the estrogen-metabolizing genes CYP17, CYP1A1, and COMT: a multigenic study on cancer susceptibility. Cancer Research, vol. 59, no. 19, pp. 4870-4875, 1999.
102. Garcia-Closas M, Glutathione S-transferase mu and theta polymorphisms and breast cancer susceptibility. Journal of the National Cancer Institute, vol. 91, no. 22, pp. 1960-1964, 1999.

103. Ishibe N, et al. Cigarette smoking, cytochrome P450 1A1 polymorphisms, and breast cancer risk in the Nurses' Health Study. Cancer Research, vol. 58, no. 4, pp. 667-671, 1998.
104. Millikan RC, et al. Cigarette smoking, N-acetyltransferases 1 and 2, and breast cancer risk. Cancer Epidemiology Biomarkers & Prevention, vol. 7, no. 5, pp. 371-378, 1998.
105. Hunter DJ, et al. A prospective study of NAT2 acetylation genotype, cigarette smoking, and risk of breast cancer. Carcinogenesis, vol. 18, no. 11, pp. 2127–2132, 1997.
106. Gertig DM, et al. N-acetyl transferase 2 genotypes, meat intake and breast cancer risk. International Journal of Cancer, vol. 80, no. 1, pp. 13-17, 1999.
107. Hines LM, et al. A prospective study of the effect of alcohol consumption and ADH3 genotype on plasma steroid hormone levels and breast cancer risk. Cancer Epidemiology Biomarkers & Prevention, vol. 9, no. 10, pp. 1099-1105, 2000.
108. Haiman CA, et al. The relationship between a polymorphism in CYP17 with plasma hormone levels and breast cancer. Cancer Research, vol. 59, no. 5, pp. 1015-1020, 1999.
109. Guillemette C, et al. Association of genetic polymorphisms in UGT1A1 with breast cancer and plasma hormone levels. Cancer Epidemiology Biomarkers & Prevention, vol. 10, no. 6, pp. 711-714, 2001.
110. Haiman CA, et al. A polymorphism in CYP17 and endometrial cancer risk. Cancer Research, vol. 61, no. 10, pp. 3955-3960, 2001.
111. Haiman CA, et al. The androgen receptor CAG repeat polymorphism and risk of breast cancer in the Nurses' Health Study. Cancer Research, vol. 62, no. 4, pp. 1045-1049, 2002.
112. Curtis, et al. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. Annals of Human Genetics, vol. 65, no. 1, pp. 95-107, 2001.
113. Seli E, Arici A. Ovulation induction with clomiphene citrate.
Available at: http://www.uptodate.com
114. Peris A. Clomiphene Therapy.
Available at: http://www.fertilitext.org/p2_doctor/clomid.html
115. Becker KL, Bilezikian JP, Bremner WJ, Hung W, Kahn CR, Loriaux DL, Nylen ES, Rebar RW, Robertson GL, Snider RH. Principles and Practice of Endocrinology and Metabolism, 3rd edition. Lippincott Williams & Wilkins, 2001.
116. Rossing MA, Daling JR, Weiss NS, Moore DE, Self SG.. Risk of breast cancer in a cohort in infertile women. Gynecol Oncol. 1996 Jan;60(1):3-7.
117. Venn A, Watson L, Bruinsma F, Giles G, Healy D. Risk of cancer after use of fertility drugs with in-vitro fertilisation. Lancet. 1999 Nov 6;354(9190):1586-90.
118. Brinton LA, Scoccia B, Moghissi KS, Westhoff CL, Althuis MD, Mabie JE, Lamb EJ. Breast cancer risk associated with ovulation-stimulating drugs. Hum Reprod. 2004 Sep;19(9):2005-13. Epub 2004 Jun 24.
119. Reich O and Regauer S. Do drugs that stimulate ovulation increase the risk for endometrial stromal sarcoma? Human Reproduction Vol.20, No.4 pp. 1112, 2005.
120. Brinton LA, Scoccia B, Moghissi KS, Westhoff CL, Althuis MD, Mabie JE, Lamb EJ.

Reply: Do drugs that stimulate ovulation increase the risk for endometrial stromal sarcoma? Human Reproduction Vol.20, No.4 pp. 1112–1113, 2005.
121. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics. 2005 Apr; 6(2):227-39.
122. Slonim N, Tishby N. Agglomerative Information Bottleneck. In NIPS'99.
123. Dougherty, J, Kohavi, R, & Sahami, M. Supervised and unsupervised discretization of continuous features. Proceedings of the Twelfth International Conference on Machine Learning, 1995. Tahoe City, CA. pp. 194—202.
124. Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, Pfeiffer RM. Cancer risk prediction models: a workshop on development, evaluation, and application. Journal of the National Cancer Institute, 2005; 97(10): 715-23.
125. National Cancer Institute. Breast Cancer Prevention Studies.
Available at: http://www.cancer.gov/cancertopics/factsheet/Prevention/breast-cancer