A Nearest Neighbor Approach to Build a Readable Risk Score for Breast Cancer

Emilien Gauthier, Laurent Brisson, Philippe Lenca, Stéphane Ragusa

A nearest neighbor approach to build a readable risk score for breast cancer

Émilien Gauthier[1,2], Laurent Brisson[1] Philippe Lenca[1], and Stéphane Ragusa[2]

[1] Institut Telecom, Telecom Bretagne, UMR CNRS 3192 Lab-STICC,
Technopôle Brest Iroise CS 83818, 29238 Brest Cedex 3, France.
(emilien.gauthier, laurent.brisson,
philippe.lenca)@telecom-bretagne.eu
[2] Statlife company, Institut Gustave Roussy,
114 rue Edouard Vaillant, 94805 Villejuif Cedex, France
(emilien.gauthier, stephane.ragusa)@statlife.fr

Abstract. According to the World Health Organization, starting from 2010, cancer has become the leading cause of death worldwide. Prevention of major cancer localizations through a quantified assessment of risk factors is a major concern in order to decrease their impact in our society. Our objective is to test the performances of a modeling method that answers to needs and constraints of end users. In this article, we follow a data mining process to build a reliable assessment tool for primary breast cancer risk. A *k*-nearest-neighbor algorithm is used to compute a risk score for different profiles from a public database. We empirically show that it is possible to achieve the same performances as logistic regressions with less attributes and a more easily readable model. The process includes the intervention of a domain expert, during an offline step of the process, who helps to select one of the numerous model variations by combining at best, physician expectations and performances. A risk score made of four parameters: *age*, *breast density*, *number of affected first degree relatives* and *breast biopsy*, is chosen. Detection performance measured with the area under the ROC curve is 0.637. A graphical user interface is presented to show how users will interact with this risk score.

1 Introduction

As cancer is becoming the leading cause of death worldwide, prevention of major types of cancer through a quantified assessment of risk is a major concern in reducing its impact in our society. Physicians have to inform patients about risk factors and have to detect fatal diseases as soon as possible in order to treat them as quickly as possible. Nowadays, this detection is led by prevention programs designed to target highest-risk subsets of the population. For example, women over 50 years old in France and over 40 in USA are recommended to perform a mammography every two years to detect breast cancer; mammography being the primary method for detecting early stage breast cancer which is the most common cause of cancer for women [18]. As a consequence, our society could benefit from a widely used risk score in order to give more accurate counseling on how cancer is impacted by risk factors and to target smallest subset of the population with higher risks. For example, using age at first mammogram as an

actionable variable, screenings programs for breast cancer could be extended: younger women with high risk profiles could be offered more frequent screenings in order to decrease death risk [27].

Even if some women may have genetic predisposition for breast cancer, environmental factors have a large impact on the risk according to Lichtenstein [22]. Because of this impact and due to acquisition cost and easyness-to-use constraints, we have decided to focus on environmental factors as attributes to compute a risk for women who never had breast cancer.

As pointed out by Testard-Vaillant [28], "*information, dialog and more patient involvement in the decision-making process*" are key words in dealing with cancer, therefore a major challenge in the field of medical counseling is to provide physicians and radiologists with adequate tools to help them to assess their patients breast cancer risk and to show easily how risk factors impact global risk. For many years, risk scores built upon statistical models were not adopted in medical counseling domain despite their performance. This may be because end-users of these tools are not oncologists nor clinicians and underlying models are too complex and too difficult to use during a medical consultation. Thus, to build a new risk score tool, we need to consider the model readability and the current medical decision process. Moreover, we will have to consider the obligation to use imbalanced datasets with missing data. To the best of our knowledge, no one has been interested in analyzing, with a mining approach, data from women who never had cancer in order to create a risk score with a prevention purpose.

Showing similar cases may improve communication with the patient, therefore increase its involvement in the prevention and decision process. Because core concept of k-nearest-neighbor algorithm is to gather similar profiles using a distance computation, we use it with help of a domain expert in order to build a tool to predict breast cancer risk and measure its performances.

The paper is organized in seven sections. Section 2 provides an overview of related works on risk models; section 3 presents our approach of the data mining process we follow; section 4 summarizes needs and constraints of users for the final tool; section 5 describes source data and section 6 reports results, discuss them and present future works.

2 Breast cancer risk scores

2.1 Statistical approaches

We present studies focusing on prevention and the use of environmental factors such as reproductive and medical history. One major risk prediction model emerges in the statistical field.

Based on an unstratified, unconditional logistic regression analysis, the most commonly used model was developed by Gail *et al* [15] using data from the *Breast Cancer Detection Demonstration Program*. Risk factor information was collected during a home interview and the analysis was based on approximately 6000 cases and controls. Among 15 risk factors obtained through patient interviews, only 5 were chosen: age, age at menarche (first natural menstrual period), number of previous breast biopsies,

age at first live birth and number of first-degree relatives with breast cancer. Gail's risk score was validated on the population of United States with the *Cancer and Steroid Hormone Study* (CASH) by Costantino *et al* [6] and in Italy on the *Florence-EPIC Cohort Study* by Decarli *et al* [8]. Chen *et al* [5] enhanced the Gail model by modeling the risk with a new equation that includes the breast density. Both regression equation parameters and coefficients are very differents than Gail's ones. It does not facilitate praticians understanding of risk evolution when adding new risk factors as attributes to describe the risk level.

Barlow *et al* [3] also built a risk prediction model using a logistic regression on the *Breast Cancer Surveillance Consortium* (BCSC) database (see Table 1 and download data from *http://breastscreening.cancer.gov*) which contains 2.4 millions screenings mammograms and associated self-administered questionnaires (see section 5). Two logistic regression risk models were built with 4 or 10 risk factors depending on the menopausal status. Compared to Gail's model, it gains the use of breast density and hormone therapy. As we will use the same database, it is worth highlighting that reported area under ROC curve (see performance measurement in section 3.4) was 0.631 for premenopausal women and 0.624 for postmenopausal women.

Primary goal of these studies was not readability, but rather highest risk detection performances and impact levels of each risk factors.

2.2 Data mining approaches and imbalanced data

Most similar data mining approaches dealt with slightly imbalanced data, mostly used to predict a cancer relapse as a result of the *Surveillance, Epidemiology and End Results* (SEER) database use. Here, we present two significant related studies involving both medical data and mining algorithm.

Endo *et al* [11] implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of the SEER program with high rate of positive examples (18.5 %). Authors did not used ROC curve to assess performances results but accuracy, specificity and sensitivity. Logistic regression had the highest accuracy, artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

Jerez-Aragonés *et al* [20] built a decision support tool for the prognosis of breast cancer relapse. They used similar attributes as Gail (like age, age at menarche or first full time pregnancy, see section 2.1) but also biological tumor descriptors. A method based on tree induction was conceived to select the most relevant prognosis factors. Selected attributes were used to predict relapse with an artificial neural network by computing a Bayes *a posteriori* probability in order to generate a prognosis system based on data from 1,035 patients of the oncology service of the Malaga Hospital in Spain .

Such studies show how mining approaches can be used to built classification tools on medical databases while dealing with missing data and business processes. But they do not consider problems (such as readability) encountered by patients who never had cancer nor physicians in their day to day interactions. Moreover, these approaches aim at predicting a class for unlabeled data (e.g. cancer relapse or not) while our goal is to provide a risk level without making the decision (breast cancer or not) in place of the

physician.

To build a risk score that helps to detect highest risk profiles among general population, the mining algorithm has to provide a risk value without labeling a woman profile. Dealing with general population means we are facing highly imbalanced data with a breast cancer incidence rate lower than 1 000 new cases for 100,000 women. Dealing with such imbalanced data can be done at both algorithmic [21] and data levels [29, 30]. At data level by choosing a different cost or rebalancing positives or negatives examples. At algorithmic level, it is possible to make a k-nearest-neighbor algorithm more sensitive to the minority class by modifying the neighborhood boundaries [21] or by using a class confidence weight [23] to handle imbalanced data during the labeling step.

3 Proposed process to build a risk score

3.1 Main objectives

The main objective of our approach is to provide physicians with a tool to assess a cancer risk level for their patient and to promote dialog between them. As statistical models spread with difficulty in the physician community, we aim to find models with good scoring performance and good readability. In our case, we say a model has a good readability if it allows a physician to explain the risk score to his patient:

- it has to be quickly readable by a physician during a medical appointment
- and has to give access to understanding the score,

Furthermore, we have other constraints: physicians have *a priori* ideas about good attributes of a model, patients need actionable attributes to change their lifestyle, both of them want immediately usable score (i.e. very low cost of data acquisition). In addition, a generic algorithm that can be easily adapted to various pathologies is desirable.

3.2 General process

Our approach follows the CRoss Industry Standard Process for Data Mining (CRISP-DM) [4] data-mining methodology. Figure 1 shows the 6 steps of this process where gray ones identify our major contributions. Business and data understanding steps are not impacted because we want to work on the same data as [3] to be able to compare our results.

Business understanding An expert with knowledge of the needs of physicians help us to prioritize our objectives (see section 3.1) and to assess the situation. We decide to focus on a scoring task (no classification or prediction).

Data understanding Despite limitations described in section 5, the BCSC database contains most of the known breast cancer personal factors. It is the largest database publicly available that includes breast density information.

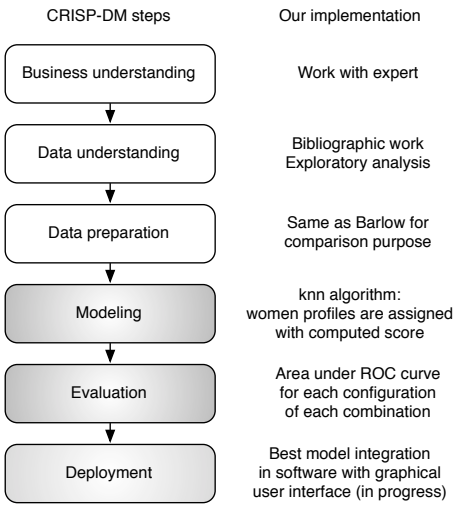

Fig. 1: General process based on CRISP-DM methodology - Gray steps identify our major contributions

Data preparation To deal with data imbalance, we can apply rebalancing algorithms on this data but it is not the focus of the paper. We do want to minimize modification of data in order to compare our results with Barlow's. The only modification we apply is normalization. It was decided to keep the same split between training and validation set.

Modeling Several data mining algorithms were considered at first, but domain expert suggested to use a k-nearest-neighbor algorithm because it uses a concept of similarity which is easily understandable by end-users without explaining a complex formula. Moreover, such algorithm is able to deal with imbalanced data if there is enough positive examples among neighbors. We generate models and search for the best combination of attributes by performing an exhaustive search (see section 3.3) on a limited set of combinations. The reason is that the expert issued a recommendation of using a restricted number of factors to make the risk score easy to use. Obviously, for large combinations, computation time can increase sharply, but it is not a problem as models are generated offline only once by us (see section 4.2), when a physician uses the final software, no computation is necessary.

Evaluation Generated models can be evaluated from a discrimination or a calibration perspective. Discrimination is needed to assess if women with breast cancer from the validation set are given higher scores than women without breast cancer. We use the Receiver Operating Characteristic method using Area Under Curve (AUC) in order to sorts models by scoring performance . Calibration is needed to assess if the number of

predicted breast cancer cases is in line with the observed number of breast cancer cases in the validation set. We use the ratio of expected cases number to observed cases number to compare models. Explanation for both evaluation criteria are given in section 3.4.

Deployment We are currently working to incorporate selected model configuration into a computer software tool for physicians. It will come with a graphical explanation of the concept of nearest neighbor. But it will not embed the database.

3.3 Focus on k-nearest-neighbor implementation

To provide experts with several interesting models, k-nearest-neighbor algorithm (see [14, 7]) is used with various size of attributes combinations (from 1 to 6 attributes), several Minkowski generalized distance measure ($p = 1$ to 5) and several k values were used (see section 6). Performance for each of hundreds of generated combinations is tested for each values of k.

We implement the k-nearest-neighbor algorithm in two steps:

- Selection of neighborhood: for a combination of attributes (e.g. *age* and *breast density*), a score value has to be computed for each combination of values (e.g. *age=5* and *breast density=3*). To compute such score value, a neighborhood has to be defined for each values combination. To determine if a profile of the database belong to the neighborhood of a combination of values, an euclidean distance is used to compute the distance between a combination of values and every single record of the training set using a normalized version of the coding values of the BCSC database. Thus, at least k of the nearest records of the database are included in the neighborhood. The neighborhood may not have always the same size because for a given group at the same distance, if k is not reached yet, all neighbors at the same distance are added to the neighborhood.
- Scoring function: the score of a combination of values, is the ratio between the number of breast cancer cases (i.e. positive examples) and the size of the neighborhood. In epidemiology, the rate of individuals having a disease in a population is called prevalence. This rate was chosen because it is well known by physicians, easily explainable to a patient and it is directly built on the number of patient diagnosed with breast cancer among patients with a similar profile.

To deal with missing data, we keep the same decision as Barlow, i.e. assign a high value when missing. It will prevent a record with a missing value to be integrated in the neighborhood.

3.4 Focus on evaluation

Mostly two kinds of evaluation are performed for epidemiological scores: discrimination and calibration. We explain why and how we use them.

Discrimination using ROC evaluation The Receiver Operating Characteristic (ROC) [10] is used to measure discrimination due to the continuous nature of our classifier: performance has to depict how positive instances are assigned with higher scores than negative ones. The ROC curve allows to measure detection performances using a moving threshold to classify examples of the validation set. Moreover, it allows direct comparison with Barlow's results and epidemiological-based scores in general.

Negative examples labeled as positive by the algorithm are called a false positives whereas positive examples labeled as positives are called true positives. The ROC curve is plotted with the false positive rate on the X axis and the true positive rate on the Y axis [13], both rates being calculated for a given threshold. It can be summarized in one number: the Area Under the ROC Curve (AUC). The area being a portion of the unit square, its value is in the [0,1] interval. The best classifier will have an AUC of 1.0 (i.e. all positive examples are assigned with higher score than negative ones) whereas an AUC of 0.5 is equivalent to random score assignment. The AUC can also be seen as the probability that randomly chosen positive and negative examples will be correctly ranked.

Calibration using E/O ratio The *E*xpected cases number to *O*bserved cases number ratio is used to measure the calibration of a model. Women from the validation set are sorted by scoring value and the validation set is split in 10 groups. In each group, the mean score is computed and converted to an expected number of cases. The sum of the 10 expected numbers of cases is then compared to the observed number of the validation set using a ratio. The best E/O ratio is 1.0, meaning that the model predict the same number of cancer cases than the actual number of cases.

To help the expert to choose the best model, each k value of each combination of attributes is assigned with an AUC and a E/O ratio value.

4 A mediation tool for physicians and patients

Providing physicians with a tool to assess a cancer risk level for their patient and promoting dialog between them, we identified constraints that arise from the users needs, we describe a solution and a we show a graphical user interface prototype that fits users needs.

4.1 Users needs and impacts on the tool

As pointed out in the introduction, the risk score is not only used to compute a risk level, but it has to be a way to promote dialog between the patient and the physician. These constraint has two majors impacts on the process that lead to the risk score construction.

First, the risk score has to be readable in how it operates. The basics of the modeling method have to be understandable by both patients and physicians: readability impacts the choice of the algorithm used to compute a score. Need of readability also impacts attributes chosen to characterize a profile. The process to build the risk score has to allow intervention from domain expert: he will choose the best combination in terms of

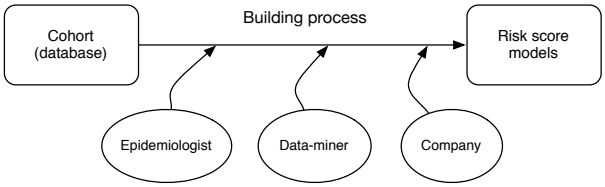

Fig. 2: Offline process with stakeholders intervention

high risk profile detection, attributes acceptability for end users and capacity to promote dialog between patient and physician, using attributes actionability for example.

Second, the risk score has to be provided in real time. To promote dialog and allow quick appropriation by users, the risk score has to be displayed instantly on a computer screen. The need of immediacy impacts the building process. The chosen algorithm has to be used in a way that allow results to be instantly available. Need of immediacy also impacts the way attributes have to be chosen depending on their time and price of acquisition. For example genetic or blood sample tests are excluded, while questions about lifestyle and women relatives are allowed.

4.2 An offline process to create the risk score

Three major constraints affect our process to build a risk score in a way that results in building our risk score in an offline manner :

- as explained in section 4.1, the risk score level has to be displayed in almost real time. Computing all profiles risk scores offline makes instant display very easy, especially when using a k-nearest-neighbor algorithm may lead to large computation time (see modeling step in section 3.2).
- very often, epidemiology databases are not publicly available because health data are sensitive and their collection are expensive. Offline computation of risk scores prevents making data available in a k-nearest-neighbor based software.
- all stakeholders have to intervene in the process of building the risk score models (see Fig. 2). Having the attributes selection and modeling steps done offline allows to implement in our process the domain expert, the contractor and the data-miner recommendations.

4.3 An online graphical user interface prototype

As all computation will be done offline, risk score values will be displayed instantly through a responsive graphical user interface (see Fig. 3).

On the graphic, the curve represents the standard incidence of breast cancer depending on the age of the woman. The curve does not evolve when using the software. At the top of the vertical line, the circle represents the woman risk: if the circle is over the

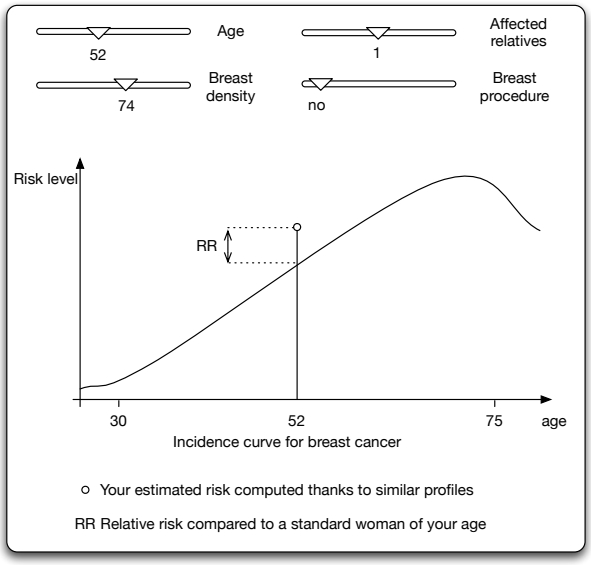

Fig. 3: A graphical user interface prototype

curve, the relative risk to standard women is over 1.0 meaning the risk is higher than the average woman of her age. If under, the relative risk is under 1.0, meaning that the risk is lower than the average woman of her age.

Each time a cursor is moved, the graphic will be instantly updated to reflect the risk of the profile. It means that the appropriation of the evolution of the risk level will be made easier for users. Patients or physicians will be able to enter the profile of a woman thanks to the sliders at the top of the interface in order to display the risk level based on real data sources used during the offline part of the process.

This kind of graphical user interface will be tested through a platform in the biggest health center dedicated to oncology in Europe, the Gustave Roussy Institute, using french data [16].

5 Data source

To build such a graphical tool and to ensure result reproducibility, we have to run the offline part of the process and therefore choose a public database with environmental factors to compute risk levels. The Breast Cancer Surveillance Consortium (BCSC) makes available a database that fits these major constraints. Each of the 2,392,998 lines match to a screening mammogram for a woman. This publicly available database provides 12 attributes to describe the woman including cancer status.

Table 1: BCSC database publicly available attributes

Full name	Short name	Description & coding
Menopausal status	menopaus	Premenopausal or postmenopausal
Age group	agegrp	10 categories from 35 to 84 years old
Breast density	density	BI-RADS breast density codes
Race	race	White, Asian/Pacific Islander, Black, Native American, Other/Mixed
Being hispanic	hispanic	Yes or no
Body mass index	bmi	4 category from 10 (underweight) to 35 and more (obese)
Age at first birth	agefirst	Before or after 30 at first live birth or nulliparous (i.e. no children)
First degree relatives	nrelbc	Number of first degree relatives with breast cancer 0, 1 or more than 2
Had breast procedure	brstproc	Prone to breast biopsy, yes or no
Last mammogramm	lastmamm	Last mammogram was negative or false positive
Surgical menopause	surgmeno	Natural or surgical menopause
Hormone therapy	hrt	Being under hormone therapy
Cancer status	cancer	Diagnosis of invasive breast cancer within one year, yes or no

5.1 BCSC database: data collection

Originally, the consortium was conceived to enhance understanding of breast cancer screening practices [2]. The consortium aims at establishing targets for mammography performance and a better understanding of how screenings affect patients in term of actions taken after the mammography. Domain experts from the surveillance consortium identified critical data elements for evaluating screenings performances reaching a consensus on a standard set of core data variables. Then, from 1996 to 2002, data were collected in seven centers across the United States: mammograms and their detailed analysis were collected and, at the same time, women were asked to complete a self-administrated questionnaire.

BCSC database provides personal factors (see Table 1) such as factual factors (age, race, body mass index), reproductive history (age at first birth, menopausal status, hormone therapy) and medical history (number of first degree relatives with breast cancer or type of menopause). In addition, breast density was recorded when the classic Breast Imaging Reporting and Data System (BI-RADS) [26] was used by the radiologist. To ensure good quality of data, exclusion rules were set: for example, women who have undergone cosmetic breast surgery were excluded as well as women with previous breast cancer and women with no known prior mammogram.

Eventually, breast cancer cases were identified by linking cancer registries to BCSC database, i.e. for each record of the database, the class of the example is positive if the corresponding women was diagnosed with breast cancer within one year after the mammogram and completing the questionnaire and negative otherwise.

Table 2: Missing data level by attribute

Attribute	Missing data level
Body mass index	55.9 %
Age at first birth	55.5 %
Surgical menopause	52.1 %
Hormone therapy	41.0 %
Breast density	26.3 %
Last mammogramm	23.4 %
Being hispanic	20.3 %
Race	15.9 %
First degree relatives	15.2 %
Had breast procedure	10.5 %
Menopausal status	7.6 %
Age group	0 %
Cancer status	0 %

Table 3: Breast cancer incidence rate per 100,000

Age category	SEER rate (2003-2007)	BCSC rate (1996-2002)
35-39	58.9	142.7
40-44	120.9	168.1
45-49	186.1	250.5
50-54	225.8	360.7
55-59	280.2	436.4
60-64	348.9	478.5
65-69	394.2	512.3
70-74	410.0	575.1
75-79	433.7	632.0
80-84	422.3	709.4
85+	339.2	Unavailable

5.2 BCSC database: exploratory analysis

Among the 2,392,998 records of the database, 9,314 cases of invasive breast cancer were diagnosed in the first year of follow up. We are facing highly imbalanced data with a positive class accounting for only 0.39 % of all records.

We also observe a high level of missing data (see table 2). Two main reasons explain missing data:

- Data were collected in different registries with non-standardized self-reported questionnaire: some questions were not asked and for any question, each woman had the possibility not to answer.
- Collection of some risk factors did not start at the same time. For example, height and weight were added later, explaining such a high rate of missing data for the body mass index.

Last, one has to notice that data of the BCSC are not representative of the USA breast cancer incidence rate (number of new cases during a specified time for a given population). Table 3 offers a comparison between the BCSC and the SEER incidence rate [1] by age categories.

Indeed, depending on data sources, the breast cancer incidence usually increase slowly from approximatively 60 to 80 years old and starts to decrease after 80 years old. But such a slower increase or decrease does not occur in the BCSC database.

6 Experimental results

6.1 Scoring performances

An experiment set was designed to test how the k-nearest-neighbor algorithm perform on the BCSC data. As one of our constraint is to build a readable risk score (see section 3.1), we select all combinations with a size s of 1 to 6 attributes among $n = 12$

Table 4: Best discrimination performances by combination size

Metrics for all combinations by size					Metric for one combination	
Size	Combi-nations	AUC Mean	AUC Std Deviation	AUC Median	Best combination (See Table 1)	AUC
1	12	0.536	0.030	0.529	agegrp	0.614
2	66	0.563	0.031	0.553	agegrp+density	0.635
3	220	0.581	0.029	0.601	agegrp+density+brstproc	0.641
4	495	0.593	0.026	0.597	agegrp+density+brstproc+lastmamm	0.642
5	792	0.602	0.023	0.586	agegrp+density+brstproc+lastmamm+menopaus	0.642
6	924	0.607	0.019	0.603	agegrp+density+brstproc+lastmamm+hrt+nrelbc	0.637

available attributes, meaning we have $\sum_{s=1}^{6} \frac{n!}{s!(n-s)!} = 2509$ combinations to test. A first way of assessing results of these combinations is to look at the best combinations by size (see Table 4). These results are obtained in an euclidian space using a 2-norm euclidian distance as they are not significantly better, when improved, using another p-norm measures.

Among one attribute combinations, *agegrp* is by far the best factor to score breast cancer risk in the BCSC database with an AUC of 0.614, while the next best attribute (not shown), *menopaus* for menopausal status, performs only at 0.563. This result confirms expert knowledge since it is widely known that age is a major breast cancer risk factor.

For combinations size from 1 to 3 attributes, mean, median and best AUC rise, whereas for sizes of 4 and 5 attributes, maximal performances level off around 0.64 with a slight decrease with 6 attributes for best combinations. It is interesting to obtain the best results using less possible attributes to improve model readability. Furthermore, our 3 attributes *agegrp, density, brstproc* combination has an AUC of 0.641 while in Barlow's results (see section 2.1), at least 4 attributes are needed to achieve an AUC of 0.631 on a subset of data that includes only premenopausal women only.

A first list of all possible combinations (from 1 to 6 attributes), is produced and sorted by performances (see Table 5-A). We observe that with an AUC of 0.642, the *agegrp, density, brstproc, lastmamm* combination perform better than the two specialized regression models obtained on pre- and postmenopausal women by [3].

6.2 Use of expert knowledge

As stated in section 3.1, besides scoring performances, our main objectives also include readability and integration of *a priori* ideas from physicians. This step of the process involves contribution from a domain expert (see section 3.2). From our domain expert point of view, when considering Table 5-A, it appears that the result of the last mammogram is a costly piece of information to obtain from women during a counseling appointment with a physician compared to performance improvement. Domain expert chooses to reduce his choices list to available combinations without *lastmamm*. Top 15 performances measures without *lastmamm* attribute are shown in Table 5-B.

Based on his domain knowledge, the expert highlights that the number of first degree relatives affected by breast cancer (*nrelbc*) is widely recognized as an important

Table 5: Top 15 performance results before and after expert advice

A. **Best combinations before expert advice**	AUC
agegrp, lastmamm, density, brstproc	**0.642**
menopaus, agegrp, lastmamm, density, brstproc	0.642
agegrp, density, brstproc	0.641
menopaus, agegrp, density, brstproc	0.641
bmi, agegrp, density, brstproc	0.640
bmi, agegrp, lastmamm, density, brstproc	0.640
agegrp, hispanic, density, brstproc	0.640
agegrp, density, brstproc, agefirst	0.639
agegrp, hispanic, lastmamm, density, brstproc	0.639
bmi, agegrp, density, brstproc, race	0.638
menopaus, agegrp, hispanic, density, brstproc	0.638
hrt, agegrp, lastmamm, density, brstproc	0.638
agegrp, density, brstproc, race	0.638
agegrp, surgmeno, lastmamm, density, brstproc	0.638
agegrp, lastmamm, density, brstproc, race	0.638

B. **Best combinations after expert advice**	AUC
agegrp, density, brstproc	0.641
menopaus, agegrp, density, brstproc	0.641
bmi, agegrp, density, brstproc	0.640
agegrp, hispanic, density, brstproc	0.640
agegrp, density, brstproc, agefirst	0.639
bmi, agegrp, density, brstproc, race	0.638
menopaus, agegrp, hispanic, density, brstproc	0.638
agegrp, density, brstproc, race	0.638
menopaus, agegrp, surgmeno, density, brstproc	0.638
agegrp, hispanic, density, brstproc, agefirst	0.638
bmi, agegrp, hispanic, density, brstproc	0.638
menopaus, agegrp, density, brstproc, agefirst	0.638
bmi, agegrp, density, brstproc, agefirst	0.637
menopaus, hrt, agegrp, density, brstproc	0.637
agegrp, density, brstproc, nrelbc	**0.637**

factor in breast cancer risk whereas other risk factor, like the body mass index (*bmi*), are not that important compared to others. According to this expert, a good candidate for our risk score would be the *agegrp, density, brstproc, nrelbc* combination with an AUC of 0.637. In addition, this performance is equivalent to the best performances of Barlow's logistic regression models (AUC of 0.624 to 0.631 depending on menopausal status). This combination uses relevant attributes for physicians according to our expert and performance loss, from 0.642 to 0.637, is acceptable. Compared to the *agegrp*, the chosen combination is a valuable performance increase. Moreover the domain expert states that the acceptability of the *agegrp, density, brstproc, nrelbc* combination by physicians, is better than the acceptability of a risk score based on *agegrp* only. It is worth highlighting that on a french database, being specifically built for breast cancer studies, the age of woman attributes only performs a 0.552 [16].

Calibration results shows that the chosen combination (*agegrp, density, brstproc, nrelbc*) has an E/O ratio of 1.01. It is better than the 1.02 E/O ratio of both top combinations *agegrp, density, brstproc* and *agegrp, lastmamm, density, brstproc* (Table 5). It is also better than the 1.02 E/O ratio of *agegrp* alone.

6.3 Performances with respect to k

In order to run a k-nearest-neighbor algorithm, the size of neighborhood has to be set. Since only k closest neighbors are used to compute the ratio healthy vs. diseased, risk score value depends on k value. If the neighborhood is too small, few breast cancer cases are included and if the neighborhood is too large, patient profiles are too different: in both cases the risk score is not reliable. For each of the 2509 combinations of attributes, we tested the scoring function with 40 values of k from 100 to 100,000.

Using, as an example, the top 15 combinations from Table 5-B, we plotted the evolution of the performance (using the AUC mean) depending on the size of the neighborhood (see Fig. 4). With an undersized neighborhood, performances are low but then, as k increases, performances increase with a maximum of 0.638. From 2500 to 8400

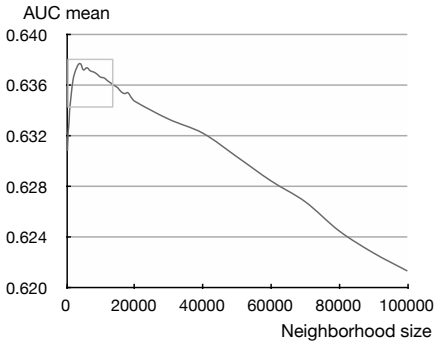

Fig. 4: Performances of top 15 combinations from Table V-B

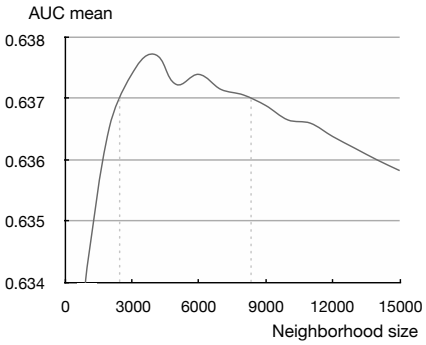

Fig. 5: Zoom on performances of top 15 combinations from Table V-B

neighbors (see Fig. 5), performances are always higher than 0.637 meaning that the algorithm is relatively stable depending on k and ultimately on the number of positive examples in the neighborhood. Eventually, as k increases, performances decrease because using a larger neighborhood leads to compute a ratio with increasingly dissimilar profiles and poor targeting.

It means that performance of the combination is not obtained with a local maximum for a single value of k. It rather depicts overall prediction ability of a combination independently of the value of k as long as the size of the neighborhood is large enough to be statistically reliable (according to the law of large numbers) and stringent enough to eliminate too dissimilar profiles.

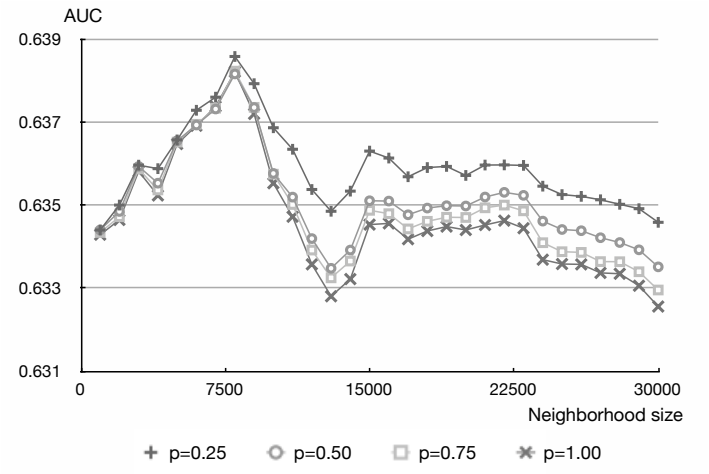

Fig. 6: Performances of best combination depending on k and the weighting function

Eliminating dissimilar profiles can be done using a weighting function [9] to decrease neighbors weight (w_i) in the prevalence computation (see section 3.3 in which w_i implicitly worth 1) depending on the euclidean distance (d_i):

$$w_i = \left(\frac{d_i}{d_{max}} \right)^p$$

with d_{max}, the greatest distance among the neighborhood and $p \in \{0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0\}$.

The prevalence is computed for the *agegrp, density, brstproc, nrelbc* combination selected by domain expert with $k \in [1000; 30,000]$ neighbors. AUC performance results are plotted in Fig. 6 only for $p = 0.25,\ p = 0.5,\ p = 0.75$ and $p = 1.0$ because performances curves for $p = 2.0, p = 3.0$ and $p = 4.0$ weighting functions are indistinguishable from curve for $p = 1.0$. Maximal performances peak is not significantly enhanced as AUC increase is less than 0.001. But when p tends to decrease, mean value of AUC increases for k in $[1000; 30,000]$. It suggests that the choice for the k value in the k-nearest-neighbor algorithm is less critical when using a weighting function because the stability range, where performances are upon a minimal value, is larger. Optimal value of k can be found more easily, making the use of the k-nearest-neighbor algorithm more independant from k.

6.4 Discussion

As statistical risk scores are not commonly used in the medical community, we think there is a possibility to improve risk scores to offer both readability in its elaboration and possibility for experts to integrate their knowledge (regarding end users expectations and the disease itself) in the process. A standard methodology called *CRISP-DM* was

followed in the process of building such a risk score. The database from the BCSC was selected because a regression-based score was already built upon it and because the database itself was publicly available. We chose to run extensive test with a k-nearest-neighbor algorithm to score profiles with different combinations of attributes. Every combinations with 1 to 6 attributes were tested, each for several values of k neighbors. Thus, we were able to allow experts to establish rules to keep or reject combinations by weighting between performance versus attributes usefulness and risk factors expected by physicians.

Nevertheless, our study has some limitations. First, on one hand, even if we selected one of the few databases large enough to be representative of the targeted population, findings from database of volunteers require cautious extrapolation to general population. On the other hand, as we use prevalence to link a profile to a risk level, even if some profiles are under or over-representated compared to general population, it has limited impact on the risk score because we used the prevalence as a score value. Second, if the concept of similarity used in the algorithm is easy to understand for everyone, performances may be limited due to imbalanced data and the constraint of not modifying data used in this paper in order to be able to compare results. However, options are available to improve steps of the process. Better performances may be obtained using another algorithm, potentially with balance of data in the data preparation step [12, 19, 24], or by combining k-nearest-neighbor with another algorithm [21, 23, 25]. Use of expert knowledge could be improved by selecting models which are provided to the expert to avoid complications due to the size of the list of combinations.

Increased acceptatbility could be reached by integrating actionable attributes. Indeed, to make more interactive softwares and increase patient involvment in the risk measurement process, actionnable risk factors as attributes may improve the prevention process with the goal to lower the risk. It implies close work with epidemiologists who lead data collection.

Since k-nearest-neighbor algorithm gives good results, we will continue to test this process on another database that include continuous attributes that were not discretized. For example age or breast density are some of the most predictive attributes and more specific data should improve performances. Higher risk profiles should be more accurately targeted leading to increased performances.

In the same time, we are developing softwares for physicians use based on prototype presented in section 4.3.

7 Conclusion

On a medical dataset, we obtain good results on readability on the modeling method with a k-nearest-neighbor algorithm easy to understand for physicians and patients. In addition, the score is very easy to use for end-users with only four attributes needed through a prototype of a graphical user interface. Thanks to our offline process, we also allow the expert to choose a combination that has not necessarily the best detection performance, but show qualities like physician acceptance and inclusion of performant attributes recognized by the community.

Our approach is innovative and successful because we have shown that it is possible to build a simple and readable risk score model for primary breast cancer prevention that performs as good as widely used logistical models.

References

1. Altekruse, S.F., Kosary, C.L., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., Ruhl, J., Howlader, N., Tatalovich, Z., Cho, H., Mariotto, A., Eisner, M., Lewis, D.R., Edwards, B.K.: SEER Cancer Statistics Review, 1975-2007 (2010)
2. Ballard-Barbash, R., Taplin, S., Yankaskas, B., Ernster, V., Rosenberg, R., Carney, P., Barlow, W., Geller, B., Kerlikowske, K., Edwards, B., Lynch, C., Urban, N., Chrvala, C., Key, C., Poplack, S., Worden, J., Kessler, L.: Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. American Journal of Roentgenology 169(4), 1001–1008 (1997)
3. Barlow, W.E., White, E., Ballard-Barbash, R., Vacek, P.M., Titus-Ernstoff, L., Carney, P.A., Tice, J.A., Buist, D.S.M., Geller, B.M., Rosenberg, R., Yankaskas, B.C., Kerlikowske, K.: Prospective breast cancer risk prediction model for women undergoing screening mammography. Journal of the National Cancer Institute 98(17), 1204–1214 (2006)
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T.: CRISP-DM 1.0 step-by-step data mining guide. Tech. rep., The CRISP-DM Consortium (2000)
5. Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C., Benichou, J., Gail, M.H.: Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. Journal of the National Cancer Institute 98(17), 1215–1226 (2006)
6. Costantino, J., Gail, M., Pee, D., Anderson, S., Redmond, C., Benichou, J., Wieand, H.: Validation studies for models projecting the risk of invasive and total breast cancer incidence. Journal of the National Cancer Institute 91(18), 1541–8 (1999)
7. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
8. Decarli, A., Calza, S., Masala, G., Specchia, C., Palli, D., Gail, M.H.: Gail model for prediction of absolute risk of invasive breast cancer: Independent evaluation in the florence-european prospective investigation into cancer and nutrition cohort. Journal of the National Cancer Institute 98(23), 1686–1693 (2006)
9. Dudani, S.A.: The distance-weighted k-nearest-neighbor rule. Systems, Man and Cybernetics, IEEE Transactions on SMC-6(4), 325 –327 (1976)
10. Egan, J.P.: Signal detection theory and ROC analysis. Series in Cognition and Perception, Academic Press (1975)
11. Endo, A., Shibata, T., Tanaka, H.: Comparison of seven algorithms to predict breast cancer survival. Biomedical Soft Computing and Human Sciences 13 2, 11–16 (2008)
12. Fan, X., Tang, K., Weise, T.: Margin-Based Over-Sampling Method for Learning From Imbalanced Datasets. In: Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, Springer (2011)
13. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
14. Fix, E., Hodges, J.L.: Discriminatory analysis, non-parametric discrimination: consistency properties. Tech. rep., USAF Scholl of aviation and medicine, Randolph Field (1951)
15. Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C., Mulvihill, J.J.: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute 81(24), 1879–1886 (1989)

16. Gauthier, E., Brisson, L., Lenca, P., Clavel-Chapelon, F., Ragusa, S.: Challenges to building a platform for a breast cancer risk score. In: Sixth International Conference on Research Challenges in Information Science. pp. 1–10. IEEE (2012)
17. Gauthier, E., Brisson, L., Lenca, P., Ragusa, S.: Breast cancer risk score: a data mining approach to improve readability. In: The International Conference on Data Mining. pp. 15–21. CSREA Press (2011)
18. IARC: World Cancer Report. IARC Publications (2008)
19. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5), 429–449 (2002)
20. Jerez-Aragonés, J.M., Gómez-Ruiz, J.A., Ramos-Jiménez, G., Muñoz-Pérez, J., E., A.C.: A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine 27, 45–63(19) (2003)
21. Li, Y., Zhang, X.: Improving k nearest neighbor with exemplar generalization for imbalanced classification. In: Huang, J., Cao, L., Srivastava, J. (eds.) Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining. Lecture Notes in Computer Science, vol. 6635, pp. 321–332. Springer Berlin / Heidelberg (2011)
22. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K.: Environmental and heritable factors in the causation of cancer, analyses of cohorts of twins from sweden, denmark, and finland. New England Journal of Medicine 343(2), 78–85 (2000)
23. Liu, W., Chawla, S.: Class confidence weighted knn algorithms for imbalanced data sets. In: Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining. Lecture Notes in Computer Science, vol. 6635, pp. 345–356. Springer (2011)
24. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B 39(2), 539–550 (2009)
25. Pham, N.K., Do, T.N., Lenca, P., Lallich, S.: Using local node information in decision trees: Coupling a local labeling rule with an off-centered entropy. In: The International Conference on Data Mining. pp. 117–123. CSREA Press (2008)
26. Reston, V. (ed.): Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). American College of Radiology (2003)
27. Teams, F.C.: Mammographic surveillance in women younger than 50 years who have a family history of breast cancer: tumour characteristics and projected effect on mortality in the prospective, single-arm, fh01 study. The Lancet Oncology 11(12), 1127–1134 (2010)
28. Testard-Vaillant, P.: The war on cancer. CNRS international magazine 17, 18–21 (2010)
29. Visa, S., Ralescu, A.: Issues in mining imbalanced data sets - a review paper. In: Sixteen Midwest Artificial Intelligence and Cognitive Science Conference. pp. 67–73 (2005)
30. Weiss, G.M., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research 19, 315–354 (2003)