Challenges to build an information system for a breast cancer risk score

Émilien Gauthier*‡, Laurent Brisson*, Philippe Lenca*, Françoise Clavel-Chapelon**, Stéphane Ragusa‡

*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

‡Statlife
Institut Gustave Roussy, 114 rue Edouard Vaillant, 94805 Villejuif Cedex, France
{emilien.gauthier || stephane.ragusa} @statlife.fr

**INSERM, Center for Research in Epidemiology and Population Health, U1018
Institut Gustave Roussy, 114 rue Edouard Vaillant, 94805 Villejuif Cedex, France
clavel@igr.fr

***Abstract*—Cancer has become the leading cause of death worldwide according to the World Health Organization. As a consequence, prevention and screening programs are set up by health authorities to decrease its incidence. Programs efficiency can be increased by targeting highest risk subset of the population. Efficient information systems capable of monitoring the population risk are thus needed. Constraints to build such a cancer risk score and their impacts on the information system are presented. As it will be shown, beyond risk score performance, a major constraint concerns the place of domain expert and the acceptability by end users. Readability then becomes an important criteria. It is shown that a simple k-nearest-neighbor algorithm can achieve good performance with the help of the domain expert. As an illustration, a risk score made of only four attributes is presented for the french population.**

I. Introduction

According to the World Health Organization, starting from 2010, cancer has become the leading cause of death worldwide [1] while, in western countries, new cases of cancer have grown at high rate in the last twenty years. Therefore, alongside clinical research to cure each cancer, prevention become a major concern in order to decrease its impact in our society. Among four existing levels of prevention, the general population is mostly interested in the two first levels. Primary prevention, that aims at avoiding occurrence of a disease and secondary prevention, that aims to diagnose and treat a disease as soon as possible, could both benefit from a widely used risk score.

Indeed, building a risk score for a given type of cancer can be a way to promote information and dialog between the medical profession and each patient [2]. With appropriate information, anyone can improve its way of life by acting on specific risk factors to decrease its global risk. Early diagnosis is another way to sharply decrease death rate. Women with high risk could benefit from a quantified risk assessment that encourage them to follow screening program recommendations.

For example, women over 50 years old in France and over 40 years old in USA are recommended to perform a mammography every two years to detect early stage breast cancer which is the first cause of cancer for women. Generalized use of risk scores could benefit every patient, first by giving them solutions to lower their risk and second by convincing them to enter national prevention programs for specific diseases.

Risk scores have to provide very good detection capabilities to identify high risk profiles among the general population. Risk scores also have to be built on readable modelling methods that can be easily understood by any patient or medical practitioner. Meeting our objectives with those constraints make the design of such a risk score challenging because it requires to deal with heavily imbalanced data (less than 2 % of breast cancer cases), using environmental factors that haven't high prediction power, with a restricted set of algorithms that meet readability criterion.

For breast cancer, the need for an efficient and readable risk score fit in a medical process currently under development by oncologists, epidemiologists and cancer treatment specialists. In this medical process, women should be able to check their own risk whether during a counseling appointment with a physician or a gynecologist, or through a dedicated internet website. Women with a high risk profile should then be encouraged to spend a day in a new generation risk clinic. There, a team of specialists will perform screenings for early stage breast cancer and determine her individual risk of developing a breast cancer by doing all tests and examinations in a single day to lower psychological impact. If needed, medical team may create a personalized plan to treat a detected cancer or to lower these risks.

This medical process is based on the ability to detect high risk french women that might need in depth counseling, but it lacks a readable and efficient risk score adjusted to french population. We suggest a method to conceive such risk score and a practical solution formalized as an information system adapted to data accessibility, computation capacity and deployment constraints. This article is organized in five

sections. Section II provides a description of constraints of the medical process and software solutions. Section III describes existing risk models for breast cancer and source data we use. Section IV presents our data mining process and performances on the database. Conclusion is in section V.

II. From medical process to information system

A. Medical process to improve prevention and screening

First proofs of generalized early screening efficacity have been gained with breast cancer. Prevention treatments for high risk women have been proven to be efficient [3]. Breast cancer research is also a field where treatments were highly improved with, as a result, a lower mortality rate. For this reasons, further screenings and prevention improvements could help to reduce mortality strongly. Goal of the medical process is to individualize both screening and prevention with help of a risk clinic, a risk clinic where high risk women will be advised to get an appointment for in depth analyze of their risks.

In order to identify women with higher risk in the general population, a risk score can be used. Two ways are considered to get women to have their risk assessed. First an internet website that any woman will be able to visit. Second, a software component integrated in the physician or gynecologist information system that provides alerts if a woman has an high risk profile depending on data available in her medical records. Both systems will advise user of the possibility to meet a team of specialists at a risk clinic.

Risk clinic have to be a decentralized structure that receive women in a care network of a given territory. Depending on the assessed risk for a woman, she may be directed to an appropriate care facility, a mammography center for example. For women with very high risks, a single appointment in a major care facility have to allow to: do a blood test for collection of clinical data, aggregate familial medical history and undergo basics screenings such as mammography. Depending on test results, the woman will get an individualized surveillance program or be sent to an appropriate care facility if a cancer is detected.

This medical process involves 6 stakeholders. The ordering party is the government that need more efficient processes to reduce mortality. The contractor is a non-profit institution in the medical field that will gather skills to ensure the ordering party that the several steps of the process will be coherent and functional. Epidemiologists (as a domain experts) and data miners have to design an efficient, easy to use and readable risk score. End users are the physician or gynecologist that will use the risk score, but also the woman that wants to assess her risk thanks to an internet website. Oncologists will lead in depth analyze for very high risk women whereas specialists will support prevention and screening for high risk profiles. At the end of the process, the beneficiaries will be the women who will get personalized advices or adapted treatments.

B. Constraints and impacts

The medical process described in II-A is based on the assessment of a women breast cancer risk. Practitioners keep total control over the advice given to the woman, but assistance can be provided through a risk score embedded in an information system that meets two constraint types.

Constraints on risk score: Risk score has to be efficient and readable to be used in a medical environment. Indeed, despite their performances, statistical breast cancer risk scores are not widely used in a prevention context. One of the reason may be that those scores are not understood by users. Users training is an explanation but all potential users will not be able to be trained in statistics. The risk score has to be based on an understandable and effective model. Readability and effectiveness impact the kind of algorithm we choose to build our risk score.

Risk scores have to allow inclusion of specific attributes for acceptability by end users. Physicians and specialists may have *a priori* ideas about good attributes of a model in a risk computation context. In the process of creating the risk score, an attribute selection step has to allow an expert, who has knowledge of users *a priori* ideas, to intervene on the choice of attributes.

Acquisition cost has to be very low. Breast cancer risk could be assessed with a better accuracy thanks to genetic risk factors (presence of BCRA1/2 gene for example) or with costly medical examination. But in front of a website or during a medical appointment, questions have to be both short and simple. Therefore, the risk score has to be based on environmental, reproductive or familial factors that are easy to answer.

Constraints on information system: List of attributes has to be flexible. Depending on the profile of the woman, some attributes may not be available or relevant to assess the risk. For example, the type of hormone replacement therapy may not be known at time of assessment. A risk score still have to be computed without the missing information. Another example is the age at menopause: if a women is not yet in menopause, age at menopause is not a relevant attribute to use, but still, risk score have to be computed. The system have to be able to deliver a risk score depending on the attributes that are available from user.

Delivering the score has to made almost in real time. For a statistically reliable risk score, model have to be based on a large enough learning set that increases computation time. But, to be able to deliver a risk level quickly, it has to be instantly available. As a consequence, the system has to precompute scores so that it can deliver a value without dealing with all the learning set each time a score is needed. Real time is a global constraint on the system, but it allows the learning step to be time consuming if needed, to increase efficiency or readability.

In epidemiology, data that allow to study cause of diseases are evolving permanently and are not publicly available. As a consequence, being able to add examples to the learning set as data are collected must be allowed by the system architecture, as well as not embedding a non-public database in a publicly available software. Having an online system that is interrogated by a client software through a web service is

a solution to avoid distribution of the database and to allow availability of data.

To summarize, our information system has to provide a risk score that is efficient and readable, usable with missing values, available in real time while allowing to update the private database used to build it.

C. Information system overview

To meet constraints described in II-B that result from the medical process described in II-A, we propose different components built to be integrated in the system. They are designed to work with the information systems of public health departments and the information systems used by physicians in their offices, see Fig. 1.

In order to design and test risk scores, we need to ease access to:

- one or multiples database servers that stores different kinds of data about people who entered an epidemiological study. They contain scanned versions of paper questionnaires that were sent to people by mail, digitized versions of medical documents that were obtained from physicians and hospitals, database that contains raw answers obtained by optical character recognition on scanned questionnaires and digitalized medical documents. Database server also contains cleaned versions of a part of the data and generated data that were deducted from people answers. Access to these database servers is restricted because they contained sensitive data about people in a hospital environment. Access to these servers depends on the data format: raw data can be reached through a document managing system or a patient data management system and database tables can be accessed directly from a statistical software.
- one or multiples computational servers. They have to provide enough power to compute all necessary data that make up our risk score and allow to run test procedures of the risk score. Useful data and executable programs have to be remotely sent on the server. Computation power can be reached through executable programs with a secured remote access.

The data miner computer acts as a platform to access all information systems that provide needed resources under various forms. From this platform, data are gathered, explored, discretized and converted into a generic format. Modelling and risk score testing decisions are automated and implemented in a software. Such software is adapted to the computational server to take advantage of its multi-core and memory capabilities. Once a risk score model is chosen, it is made available through a database table that contains precomputed risk score levels. Indeed, to meet constraints of section II-B, we choose to allow its access through an internet solution.

Users (women, physicians, specialists) will be able to access the risk score through a web service from a web server. Access will be made easy: women will be able to calculate their risk score with a website whereas physicians or gynecologists will likely use it through their information systems thanks

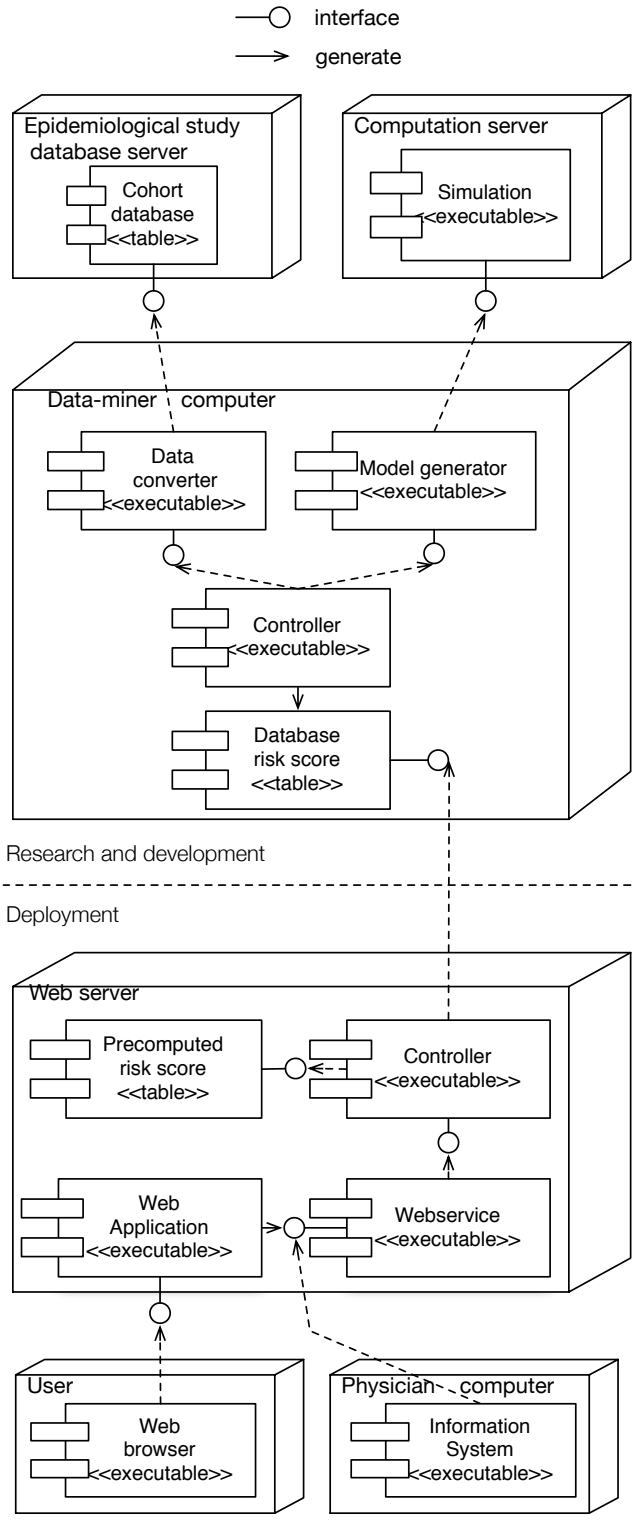

Fig. 1. Applicative architecture of information system

to components compatible with their medical software. To protect data, only precomputed risk score are uploaded on the web server that will provide a web service. It will be used to power a website for women and components that will be integrated into the different kinds of information systems owned by physicians.

III. Breast cancer: available risk scores and data

A. Breast cancer risk scores

1) Epidemiological approach: Thanks to the *Breast Cancer Detection Demonstration Program*, using an unconditional logistic regression model, Gail *et al* [4] conceived what is now the most commonly used model to predict breast cancer risk with environmental factors. Risk factors information was collected through a home interview from 6,000 women. Eventually, among 15 risk factor, only 5 were chosen to be part of the final model: age of patient, age at menarche (first natural menstrual period), age at first live birth, number of previous breast biopsies and number of first degree relatives affected by breast cancer. Cumulative risk of breast cancer is calculated by a multiplication of each of the five relative risks. Then, individual risk of a women can be computed by multiplying the cumulative risk by an adjusted population risk of breast cancer. The model was validated and adapted on several other population: asian and pacific islander american women [5], italian women [6] and on american women [7] of the CASH study (see below). On the 82,109 women of the Nurses Health Study, using Gail's model modified by Costantino [7], Rockhill reports discrimination with an AUC (Area Under ROC Curve, see performance measurement in section IV-C) of 0.58 [8].

Another well known risk score is the Claus model [9], also known as CASH model, based on the *Cancer and Steroid Hormone Study* lead by the *Centers for Disease Control and Prevention*. This genetic-based model was built using a segregation analysis for women who have a familial history of breast cancer. It aimed at understanding the breast cancer transmission model using familial data (mother and sister) of 9,418 women. Risk factors were: age of patient, number of first and second degree relatives affected by breast cancer. Individual risk can be easily obtained by reading a dedicated table depending on the number of affected relatives.

Using environmental factors, Barlow *et al* [10] proposed a risk prediction model based on the *Breast Cancer Surveillance Consortium* (BCSC) database which contains 2.4 millions screenings mammograms and answers to associated self-administered questionnaires. Using two logistic regression models, a risk score was built with 4 or 10 risk factors depending on the menopausal status. Unlike the Claus model, it can be used for women without familial history of breast cancer. Compared to Gail's, it gains use of two attributes with higher prediction power: hormonal therapy and breast density. Barlow *et al* report AUC: 0.631 for premenopausal women based model and 0.624 for postmenopausal women.

2) Data mining approach: Because most approaches of breast cancer risk prediction deal with cancer relapse in the data mining field, authors did not faced as imbalanced data as epidemiological studies did (see III-A1). However, it is worth highlighting two significant studies involving mining algorithms and medical data.

Jerez-Aragonèz *et al* [11] chose the prognosis of breast cancer relapse to build a decision support tool. A database that gather information about 1,035 patients of the Malaga Hospital, Spain was used. Similar attributes than Gail (age, age at menarche and first full term pregnancy) were analyzed along biological tumor descriptors. Selection of most relevant prognosis factor was done with a tree induction based method. Attributes were then used to predict cancer relapse with an artificial neural network by computing a Bayes *a posteriori* probability in order to generate the prognosis system.

The *Surveillance, Epidemiology and End Results* (SEER) database was used by Endo *et al* [12] to implement common machine learning algorithms to predict survival rate for women affected by breast cancer. Even if positive examples were highly represented (18,5 %) in the database and area under the ROC curve wasn't chosen as performance metric, it is relevant to note that logistic regression had the highest accuracy, artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

Because we want to predict breast cancer risk for women among the general population, we face highly imbalanced data: new breast cancer cases rate is lower than 1 %. Usually, dealing with imbalanced data can be done both at algorithmic and data level according to Japkowicz *et al* [13] and Visa *et al* [14]. Indeed, guiding the data mining process or sampling data are solutions to increase detection performances for high risk profiles, but we will not use them in our approach due to the nature of the algorithm we chose.

Jerez-Aragonèz's [11] and Endo's [12] studies show how mining techniques can be used to build classification tools on medical databases while considering missing data and processes. But, even for epidemiological risk scores, they do not consider all of our constraints, such as readability, easy of use for patients, physicians and specialized doctors in their day to day interactions and adaptation to the french population.

B. Available data

1) E3N cohort: To build a risk score for french women, we use data of one of the largest epidemiological cohort (a group of people with a shared characteristic) study in France: E3N ("Étude Épidémiologique des femmes de l'Éducation Nationale") [15]. It is a prospective study that includes women of the MGEN, "Mutuelle Générale de l'Éducation Nationale", a health insurance plan primarily covering employees of the french national education system. The E3N study is the french component of EPIC, the European Prospective Investigation into Cancer and Nutrition. Since 1990, 98,995 volunteers women, born from 1925 to 1950, were asked to fill 10 self-administered questionnaires about their lifestyle (for example diet, reproductive factors, alcohol and tobacco consumption, physical activity, etc), regular use of medical treatments (hormonal treatments for example) and personal medical history

TABLE I
DATASET ATTRIBUTES BUILT FROM THE E3N DATABASE COHORT

Full name	Short name	Description & coding
Age in 1997	*age*	round to year from 46 to 72 years old
First full term pregnancy	*fftp*	0-20;21-25;26-30;30+ years old
Number of children	*nbchild*	0;1;2;3;4+
Tobacco status	*tobbaco*	None, former or current. Occasional, regular, unknown
Body mass index	*bmi*	4 categories : 0-19;20-24;25-29;30+
First degree relatives	*kdeg1*	First degree relatives with breast cancer: 0 to 4
Breast feeding	*bfeed*	In months: 0;1-2;3-4;5+
Abortion	*abort*	Number of abortions: 0;1;2+
Age at menopause	*agemeno*	round to year from 40 to 70 years old
Menopause type	*typemeno*	Natural, artificial, never had period
Menarche	*menarche*	Age at first menstrual period
Hormone therapy	*hrt*	No, yes and type of hormone therapy
Biopsy	*biopsy*	Number of breast procedures
Alcool	*alcool*	Quantity of alcool per day per 5g
Cancer status	*cancer*	Diagnosis of invasive breast cancer within five year, yes or no

(cancer, but also cardiovascular diseases, osteoporosis, cognitive decline or diabetes). Approximately 50 to 200 questions are asked in each of the 10 questionnaires, excluding two extensive dietary questionnaires with 1,000 questions in each one of them.

As E3N is a prospective study, information was collected before any major pathology occurred, so women did not have cancer at time of inclusion in the cohort. When cancer cases are reported, invasive breast cancer cases are ascertained by obtaining histopathology reports. Up to 2011, invasive breast cancer cases were ascertained at a 92.4 % rate. Deaths information and causes of death, that may include unreported cancer cases, are obtained thanks to family members and the MGEN insurance database in accordance with the french data protection legislation.

2) Dataset construction: To conceive a breast cancer risk score that meets constraints of section II-B, we need a dataset with environmental data as attributes and cancer status as class. Several types of breast cancer exist, we define a women as being affected by breast cancer only if it is ascertained as an invasive breast cancer with no distinction depending on its hormone status. Among all attributes available from the E3N study, we decide to include in our dataset, only known risk factors. Risk factors are attributes for which we know from the literature, that epidemiologically speaking, they have an impact on the breast cancer risk. Impact should be understand at different levels: direct impact (*e.g.* hormone treatment), intermediate impact (*e.g.* age at menarche for hormone exposure), risk marker (*e.g.* number of biopsies). These 12 attributes are sum up in the Table I.

We decide to build a 5-years risk score in order to aggregate enough cancer cases and to offer a medium-term perspective to the woman and the physician. The dataset is built with attributes that describe a woman at time of answer of the fifth questionnaire in 1997: it is our baseline. As each women had several months to answer the 1997's questionnaire, each 5-years timeframe does not start at the same time for each one of them. For each woman, cancer status is defined as positive if breast cancer occurs in a 5-years delay after her 1997 response, negative if not.

Dataset includes 92,078 women. During the 5-years timeframe, 1,647 women (1.79 %) have developed invasive breast cancer: they are labeled as positive examples whereas other 98.21 % are labeled as negative examples. Dataset is therefore strongly imbalanced. For some of the risk factors, questions were asked in previous questionnaires. Those questionnaires are used to retrieve information that is not available from the 1997's questionnaire: for example age at menarche is usually between 8 and 18 years old and it does not matter if information was collected 5 years before baseline, in 1992 instead of 1997. Remaining missing values, that could not be found in any questionnaire, are replaced by median value for the attribute. Discretization was realized in order to maintain distribution of values.

IV. BUILDING OUR RISK SCORE

A. CRISP-DM based process

1) Main objectives: The main objective of our approach is to build a method to conceive a risk score and a practical solution formalized as an information system that will be used through a business oriented software by a physician or through an internet website by any woman. As statistical models spread with difficulty in the medical field, we aim to find a model with good scoring performance and good readability. We say a model has a good readability if it can be understood thanks to a simple picture or sentence: it has to be quickly readable during a medical appointment in a primary prevention context.

Furthermore, we have other constraints: patients need actionable attributes to change their lifestyle in a prevention context, physicians have *a priori* ideas about good attributes of a model in a risk computation context, both of them want immediately usable score (see II-B). In addition, a generic algorithm that can be easily adapted to various pathologies in different countries is desirable.

2) General process: Our approach follows the CRoss Industry Standard Process for Data Mining (CRISP-DM) [16] data-mining methodology. Figure 2 shows the 6 steps of this process where gray ones identify our major contributions.

Business understanding: An expert with knowledge of the needs of users help us to prioritize our objectives (see section IV-A1) and to assess the situation. We decide to focus on a scoring task (no classification or prediction).

Data understanding: The E3N cohort (see section III-B) contains numerous known breast cancer personal risk factors.

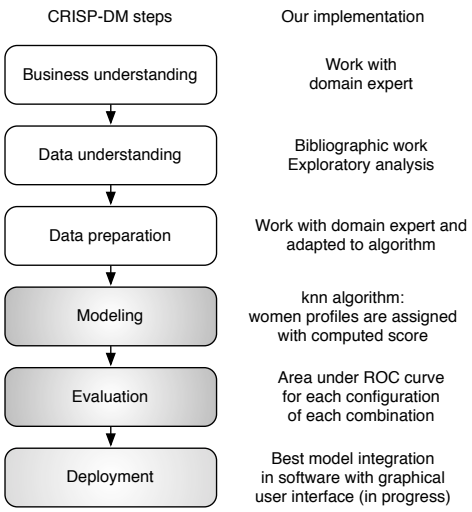

Fig. 2. General process based on CRISP-DM methodology - Gray steps identify our major contributions

Data are used by epidemiologists to discover associations between factor and cancer.

Data preparation: To deal with highly imbalanced data, we can apply rebalancing algorithms on this data but it is not the focus of the paper. As our modeling step will be based on the negative/positive ratio, we want to minimize modification of data balance in order to provide risk score closest as possible as reality. The only modification we apply is normalization thanks to a division by the standard deviation. To limit overfitting, it was decided to apply a random 75/25% split between training and validation set.

Modeling: If several data mining algorithms were considered, 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 IV-B) 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. As an example, for data described in section III-B1 and for 1 to 6 attributes combinations, it takes approximately 15 hours to compute.

The k-nearest-neighbor algorithm also meets the genericity criterion: as a non parametrical model, it fully relies on data (no statistical parameters optimization) so it can be adapted for other pathologies and other dataset from different countries by replacing the dataset.

Evaluation: We evaluate generated models with Receiver Operating Characteristic validation (see section IV-C) using Area Under Curve (AUC) in order to sorts models by scoring performance. Then, our expert has to choose the most useful models leveraging on the AUC performance combined with its knowledge of users needs. ROC evaluation of every generated model is automatized in our software as well as descriptive statistics to characterize the neighborhood: mean, median and standard deviation for both k value and number of positive examples. We are still improving our process to formalize and support expert choice through a dedicated graphical user interface. It provides useful information about each model and ease the browsing among hundreds of models by offering sorting and filtering options.

Deployment: We are currently working to develop web services that will power a web application and components for physicians information systems.

B. Focus on k-nearest-neighbor implementation

To provide experts with interesting models, k-nearest-neighbor algorithm (see [17], [18]) is used with various size of attributes combinations (from 1 to 5 attributes) and several k values were used (see section V). Performance of each of hundreds 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 *number of first degree relatives with breast cancer*), a score value has to be computed for each combination of values (e.g. *age=54* and *number of relatives with breast cancer = 1*). 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 value and every single record of the dataset using a normalized version of the values. 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 ratio of individuals having a disease in a population is called prevalence. This ratio was chosen because it is well known by physicians, easily explainable to a risk score user and it is directly built on the number of patient diagnosed with breast cancer among patients with a similar profile.

C. Focus on ROC evaluation

Our performance metric has to depict how positive instances are assigned with higher scores than negative ones: we used the Receiver Operating Characteristic (ROC) [19] to measure performance due to the continuous nature of our classifier. The

ROC curve enable to measure detection performances using a moving threshold to classify examples of the validation set. Moreover, it allows direct comparison with epidemiological-based scores from the literature.

ROC mechanism: 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 [20], both rates are calculated for a given threshold. ROC curve can be summarized in one number: the Area Under the ROC Curve (AUC). The area is a portion of the unit square, its value is in an [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.

Each k value of each tested combination of attributes is assigned with a ROC curve and the corresponding AUC in order to help the expert to choose the best model.

V. Experimental results

On a publicly available database [10], we have empirically shown that a data mining approach can provide better performances from a discrimination and a readability perspective in [21]. In this section, we report the results obtained with a dataset built on the E3N cohort with specific objective: build a risk score for french women that will be used in a global medical process called a risk clinic.

A. Expert knowledge limits dataset

Before testing performances of our k-nearest-neighbor ($k$nn) implementation, we submit our list of 12 attributes (see section III-B2) to a domain expert. With his knowledge of physicians and specialists *a priori* ideas about a good composition for a breast cancer risk score (a constraint of section II-B), he advices us to keep only 8 attributes among 12 (see Table II). He chooses this 8 attributes, or a subset of them, because they are commonly used by physicians to roughly assess their patients risk and because their are recognized in the community to be good risk factor for breast cancer assessment.

TABLE II
ATTRIBUTES CHOSEN BY DOMAIN EXPERT

Age	(*age*)	First pregnancy	(*fftp*)
Age at menopause	(*agemeno*)	Menopause type	(*menotype*)
Menarche	(*menarche*)	Hormone therapy	(*hrt*)
Biopsy	(*biopsy*)	First degree relatives	(*kdeg1*)

The first list of attributes was objectively conceived with knowledge of impact on cancer risk as only criterion. It did not depend on constraints set by process stakeholders. However, this second list is created to meet a specific constraint of the process. It is build subjectively to take into consideration, the knowledge that we have from users for the specific purpose of building a breast cancer risk for a defined set of users. We call this 8 attributes dataset, the restricted dataset.

B. Scoring performances

An experiment set was designed to test how the $k$nn algorithm performs on the restricted dataset build from the E3N cohort. As one of our constraints is to build a readable risk score, we decide to limit the size of the combination to 5 attributes. We select all combinations with a size s of 1 to 5 attributes among $n = 8$ attributes of the restricted dataset, meaning we have $\sum_{s=1}^{5} \frac{n!}{s!(n-s)!} = 218$ combinations to test. Each one of the 218 combinations was tested with 39 values of k nearest neighbors, meaning 8,502 configurations were computed.

TABLE III
DESCRIPTIVE STATISTICS BY COMBINATION SIZE

Size	Combinations	AUC Mean	AUC Std Dev.	AUC Median
1	8	0.533	0.023	0.530
2	28	0.556	0.022	0.553
3	56	0.569	0.019	0.565
4	70	0.577	0.017	0.579
5	56	0.584	0.016	0.589

Table III provides descriptive statistics by combination size. Each attributes combination appears only once in the statistics with the best AUC computed among all values of k that were tested. As the size of combinations increases, AUC increases more slowly. Decline of the standard deviation shows that combination tends to have more similar performances because combinations tends to include more risk factors with prediction power.

TABLE IV
BEST COMBINATION BY SIZE

Size	Combination	k value	AUC
1	*hrt*	3,000	0.572
2	*hrt, age*	1,800	0.593
3	*hrt, fftp, kdeg1*	10,500	0.601
4	*hrt, age, kdeg1, fftp*	9,000	0.604
5	*hrt, age, kdeg1, fftp, agemeno*	8,500	0.605

Among combinations of one attribute (see Table IV), the *hrt* combination is the best factor to predict breast cancer with an AUC of 0.572. Next best attribute is *age* with an AUC of 0.552. Usually, age is the best predictor for breast cancer risk but, as the hormone replacement treatment is given after the menopause, *hrt* attributes carry a partial information about age of the woman plus the kind of hormone replacement treatment used by women of the cohort. Epidemiologically speaking, domain expert states that this combination of one attribute has no sense and performance is not good enough, so we consider other available combinations.

Among tested combinations, the best performance is achieved by the *hrt, age, kdeg1, fftp, agemeno* combination. The attribute *agemeno* only slightly increase performances, then to meet our readability constraint, we choose not to use it. Same reasoning could apply to the 4-attributes combination compared to the 3-attributes combination, but the 3-attributes combination does not display use of *age*. Acceptability constraint would not be met without *age* attribute according to our

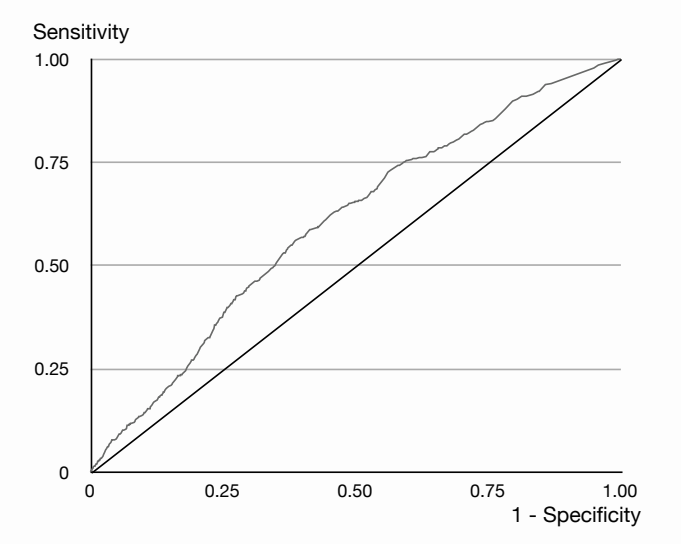

Fig. 3. ROC curve for the *hrt, age, kdeg1, fftp* combination

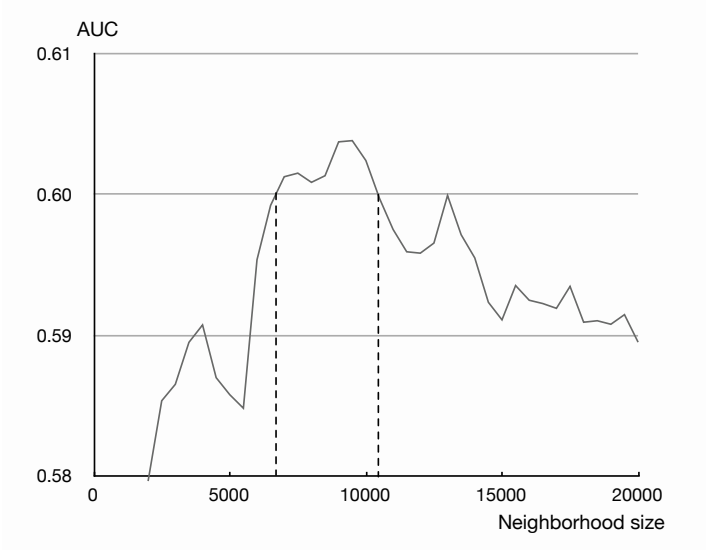

Fig. 4. AUC performance for the *hrt, age, kdeg1, fftp* combination

domain expert. Logically, the *hrt, age, kdeg1, fftp* combination is chosen.

C. Details on the chosen combination

Combination *hrt, age, kdeg1, fftp* is chosen because it meets our readability and acceptability constraints with a high level of performances compared to other combinations. Figure 3 shows ROC curve for the combination.

Neighborhood characteristics: To assign a risk score to examples of the validation set, all scores were computed on the learning set first by defining a neighborhood for each profile (the profile *age=52*, *hrt=0*, *fftp=23* and *kdeg1=1* for example), and then by computing the prevalence value (see section IV-B) for the profile. For this chosen combination, the generated neighborhoods contain at least 9,000 neighbors and a maximum 12,359 of neighbors (because k is not a strict value, see section IV-B). On average, it contains 9,296.7 neighbors (median: 9,161, standard deviation: 373,2). As the score is based on the number of breast cancer cases among neighbors, it is interesting to look at the number of cancer cases in the neighborhoods: at least 110 and a maximum of 288 cancer cases in the neighborhood, on average 235.0 cases (median: 251 and standard deviation: 38.0).

Performance stability: In order to run a $k$nn 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. Figure 4 presents evolution of the AUC depending on the size of the neighborhood.

With an undersized neighborhood, performances are low but then, as k increases, performances increase with a maximum of 0.604. From 6,900 to 10,300 neighbors, performances are always higher than 0.600 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 gathering a larger neighborhood leads to compute a ratio with increasingly dissimilar profiles.

*Enhance performance:*To increase performances, we have tried to add enhancements to our $k$nn implementation. Using Minkowski distances, as $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$ with $p = 3$ to 6 instead of an euclidian distance with $p = 2$, has not provided better AUC measures on the chosen combination. Adding the distance-weighted $k$nn rule described by Dudani [22] do not increase performances with the weighting function that were tested on the chosen combination. We decide not to try further as we do not want to overlearn.

D. Discussion

Our intention is to build an IS that integrate: access to cohort database and a computational server to conceive of a risk score that will be used to detect women with high risk for breast cancer. To achieve this purpose, a major step is to build a risk score that fit in the following medical process: risk assessment, cancer related advices, risk clinic appointment.

By testing every combination of attributes from a restricted dataset build with help of a domain expert and by carefully selecting a set of attributes depending on its discrimination capacity and the nature of the risk factors to meet our constraints, we choose the *hrt, age, kdeg1, fftp* combination. With an AUC of 0.604, the risk score have similar performances than other scores that use similar risk factors as Rockhill *et al* [8] with Gail's model on an american population. But performances are not as high as those obtained by Barlow *et al* [10] or by Gauthier *et al* [21] on the american *Breast Cancer Surveillance Consortium* database. Domain expert explains this difference by the absence of attributes with higher prediction power such as breast density. This information being currently retrieved

from women of the E3N cohort, we hope to increase performances of our risk score.

Discrimination performances may be increased using another $k$nn enhancements than those we tested, but we might prefer to focus on improving another aspects of our IS. For example, use of the software to generate combinations and test them could benefit from tracability and reproductibility improvements in order to keep track of dataset, $k$nn settings and associated results. Graphical user interface for end users also need work to ease the use of the risk score and make the $k$nn concept as readable as possible.

Nevertheless, our study has some limitations. First, even if we used one of the few databases large enough to be representative of the targeted population, findings from a database based on employees of the national education system require cautious extrapolation to general population. Conversely, as our risk score is build on a ratio between number of cancer cases and size of neighborhood, it depicts association between a woman profile ant a breast cancer risk. Over- or underrepresentation of profile in the database, compared to general population, should have limited impact on extrapolation of the built risk score. Second, use of expert knowledge could be improved: process followed by domain expert to select useful combinations could benefit from adapted graphical user interface to browse output of our $k$nn implementation.

VI. Conclusion

In order to create an efficient and readable breast cancer risk score for french women that fit in a medical process built to detect high risk women that may benefit from in depth counseling and regular screenings, we studied some constraints that have to be considered to design an information system to build risk scores for major pathologies. We then instantiate the system with a knn algorithm and the domain expert knowledge..

Discrimination performances of the $k$nn algorithm for the combination chosen by domain expert are in the same range than measures reported by studies on different population with similar attributes thanks to widely used statistical methods. Moreover our proposal meets deployment and readability constraints that make the challenge to produce a practical and readable solution usable by physicians and women.

Acknowledgment

The authors are indebted to all participants for providing the data used in the E3N study and to practitioners for providing pathology reports. They are grateful to Mrs. R. Chaït, Mrs. M. Fangon, Mrs. L. Hoang, Mrs. C. Kernaleguen and Mrs. M. Niravong for their technical assistance.

References

[1] IARC, "World Cancer Report," p. 512, 2008. [Online]. Available: http://www.iarc.fr/en/publications/pdfs-online/wcr/index.php

[2] P. Testard-Vaillant, "The war on cancer," *CNRS international magazine*, vol. 17, pp. 18–21, 2010.

[3] B. Fisher, J. P. Costantino, D. L. Wickerham, C. K. Redmond, M. Kavanah, W. M. Cronin, V. Vogel, A. Robidoux, N. Dimitrov, J. Atkins, M. Daly, S. Wieand, E. Tan-Chiu, L. Ford, N. Wolmark, other National Surgical Adjuvant Breast, and B. P. Investigators, "Tamoxifen for prevention of breast cancer: Report of the national surgical adjuvant breast and bowel project p-1 study," *J. Natl. Cancer Inst.*, vol. 90, no. 18, pp. 1371–1388, 1998.

[4] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *J. Natl. Cancer Inst.*, vol. 81, no. 24, pp. 1879–1886, 1989.

[5] R. K. Matsuno, J. P. Costantino, R. G. Ziegler, G. L. Anderson, H. Li, D. Pee, and M. H. Gail, "Projecting individualized absolute invasive breast cancer risk in asian and pacific islander american women," *J. Natl. Cancer Inst.*, 2011.

[6] A. Decarli, S. Calza, G. Masala, C. Specchia, D. Palli, and M. H. Gail, "Gail model for prediction of absolute risk of invasive breast cancer: Independent evaluation in the florence-european prospective investigation into cancer and nutrition cohort," *J. Natl. Cancer Inst.*, vol. 98, no. 23, pp. 1686–1693, 2006.

[7] J. Costantino, M. Gail, D. Pee, S. Anderson, C. Redmond, J. Benichou, and H. Wieand, "Validation studies for models projecting the risk of invasive and total breast cancer incidence." *J. Natl. Cancer Inst.*, vol. 91, no. 18, pp. 1541–8, 1999.

[8] B. Rockhill, D. Spiegelman, C. Byrne, D. J. Hunter, and G. A. Colditz, "Validation of the gail et al. model of breast cancer risk prediction and implications for chemoprevention," *J. Natl. Cancer Inst.*, vol. 93, no. 5, pp. 358–366, 2001. [Online]. Available: http://jnci.oxfordjournals.org/content/93/5/358.abstract

[9] E. B. Claus, N. Risch, and W. D. Thompson, "Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction," *Cancer*, vol. 73, no. 3, pp. 643–651, 1994.

[10] W. E. Barlow, E. White, R. Ballard-Barbash, P. M. Vacek, L. Titus-Ernstoff, P. A. Carney, J. A. Tice, D. S. M. Buist, B. M. Geller, R. Rosenberg, B. C. Yankaskas, and K. Kerlikowske, "Prospective breast cancer risk prediction model for women undergoing screening mammography," *J. Natl. Cancer Inst.*, vol. 98, no. 17, pp. 1204–1214, 2006.

[11] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and A.-C. E., "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial Intelligence in Medicine*, vol. 27, pp. 45–63(19), jan 2003.

[12] A. Endo, T. Shibata, and H. Tanaka, "Comparison of seven algorithms to predict breast cancer survival," *Biomedical Soft Computing and Human Sciences*, vol. 13 2, pp. 11–16, 2008.

[13] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[14] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets - a review paper," in *Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*, 2005, pp. 67–73.

[15] F. Clavel-Chapelon and E. Group, "E3n, a french cohort study on cancer risk factors," *European Journal of Cancer Prevention*, vol. 6, no. 473-478, 1997.

[16] P. Chapman, J. Clinton, R. Kerber, and T. Khabaza, "Crisp-dm 1.0 step-by-step data mining guide," The CRISP-DM Consortium, Tech. Rep., 2000.

[17] E. Fix and J. Hodges, "Discriminatory analysis, non-parametric discrimination: consistency properties," USAF Scholl of aviation and medicine, Randolph Field, Tech. Rep., 1951.

[18] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[19] J. P. Egan, *Signal detection theory and ROC analysis*, ser. Series in Cognition and Perception. Academic Press, 1975.

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[21] E. Gauthier, L. Brisson, P. Lenca, and S. Ragusa, "Breast cancer risk score: a data mining approach to improve readability," in *The International Conference on Data Mining*, 2011, pp. 15–21.

[22] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-6, no. 4, pp. 325 –327, april 1976.