JOURNAL OF INSURANCE MEDICINE

J Insur Med 2023;50:36–48

ADVANCES IN RISK ASSESSMENT

Methods for Inclusive Underwriting of Breast Cancer Risk with Machine Learning and Innovative Algorithms

Manuel Plisson, PhD[1]; Antoine Moll, MSc[1]; Valentine Sarrazin, MSc[1]; Denis Charles, MSc[1,2]; Thibault Antoine, MSc[1]; Razvan Ionescu, MSc[1]; Odile Koehren, BSc[1]; Eric Raymond, MD, PhD[1,2,3]

Introduction.—Due to early detection and improved therapies, the prevalence of long-term breast cancer survivors is increasing. This has increased the need for more inclusive underwriting in individuals with a history of breast cancer. Herein, we developed a method using algorithms aiming facilitating the underwriting of multiple parameters in breast cancer survivors.

Methods.—Variables and data were extracted from the SEER database and analyzed using 4 different machine learning based algorithms (Logistic Regression, GA2M, Random Forest, and XGBoost) that were compared with Kaplan Meier survival estimates. The performances of these algorithms have been compared with multiple metrics (Log Loss, AUC, and SMR). *In situ* (non-invasive) and metastatic breast cancer were excluded from this analysis.

Results.—Parameters included the pathological subtype, pTNM staging (T: tumor size, N; number of nodes; M presence or absence of metastases), Scarff-Bloom-Richardson grading, the expression of estrogen and progesterone hormone receptors were selected to predict the individual outcome at any time point from diagnosis. While all models had identical performance in terms of statistical metrics (AUC, Log Loss, and SMR), the logistic regression was the one and only model that respects all business constraints and was intelligible for medical and underwriting users.

Conclusion.—This study provides insight to develop algorithms to set underwriter-friendly calculators for more accurate risk estimations that can be used to rationalize insurance pricing for breast cancer survivors. This study supports the development of a more inclusive underwriting, based on models that can encompass the heterogeneity of several malignancies such as breast cancer.

Address for Correspondence: Prof Eric Raymond — Chief Medical Officer of Inclusive UW and Medical Expertise | Oncology – SCOR Global Life, Biometric Modelling & Inclusive Underwriting, 5 Avenue Kléber, 75795 Paris Cedex 16, France; ph: +336 7629 6115; eraymond@scor.com

Key words: Breast cancer, cancer survivors, conditional survival, machine learning, underwriting, multivariate analysis, algorithms

Author Affiliations: [1]SCOR Life & Health, Knowledge Team, 5 Avenue Kléber, 75795 Paris Cedex 16, France; [2]Université de Poitiers, CRIEF; [3]Department of Oncology, Groupe Hospitalier Paris Saint Joseph, 185 Rue Raymond Losserand, 75014 Paris, France.

Received: July 15, 2022

Accepted: March 21, 2023

According to the global estimate of the World Health Organization,[1] 2.3 million women have been diagnosed with breast cancer with 685,000 deaths in 2020. As of the end of 2020, 7.8 million women who were diagnosed are alive with a medical history of breast cancer in the past 5 years, making it the world's most prevalent cancer.[2] While the increased incidence is poorly explained, the improvement of survival that has been

observed since the 1980s are primarily explained by early detection programs combined with improved multimodality cancer treatments.[3]

As the number of breast cancer survivors is increasing, it is now estimated that 1 among 8 women are likely to develop breast cancer over their lifetime. Survivorship post-cancer programs including prevention, medical follow up, psychological support, professional, and social reinsertion have been developed to restore normal life in breast cancer survivors.[3] Many groups across the world have been starting to advocate for changing the social perception of cancer survivors, including a more inclusive approach to access insurance coverage such as for real estate loans.[4] From the insurance standpoint, breast cancer stands as a highly heterogenous tumor compared to many other malignancies with multiple parameters used to estimate the risk of relapse and survival at diagnosis.[5] This high heterogeneity increases the complexity of survival risk estimation and appears particularly important in insurance to ensure data-driven risk estimation and to justify ratings in cases of disagreement with clients and trial.[6]

At the time of diagnosis, the pathological subtypes, TNM staging (T: tumor size, N; number of nodes; M presence or absence of metastases), Scarff-Bloom-Richardson grading, the expression of estrogen and progesterone hormone receptors, HER2-neu overexpression, the Ki67 rate or mitotic index, genomic signatures, the presence of microvascular or perineural invasions have been used for several years by physicians in multivariate analyses to estimate patient prognosis and prescribe appropriate therapies.[7] To allow a better understanding of the TNM classification, it is important to remember than *c*TNM account for the *clinical* evaluation performed using physical and radiological exams, *p*TNM account for the classification performed using *pathological* specimens obtain from surgical resection, and *yp*TNM account for *post-operative pathological specimens* obtain from surgery after neo-adjuvant chemotherapy. For several years, multivariate algorithms have been routinely used by physicians through web-based calculator interfaces using machine learning.[8] For example, the UK Breast Cancer Predict web-based algorithm (https://breast.predict.nhs.uk/) was used to estimate patient risk and justify the use of systemic therapies such as chemotherapy, hormonotherapy, and monoclonal antibodies such as trastuzumab, a monoclonal antibody directed against HER2-neu receptor and restricted to patients with HER2 2+ with positive FISH and HER2 3+ cancer cells.

In the medical literature focusing on breast cancer patient outcome, it has been demonstrated that the risk of relapse and death varies over time.[9] Conditional survival was shown to be a useful concept to estimate the probability of surviving further years, given that a patient has already survived for several years after the diagnosis of a chronic disease, such as malignancies where a large number of patients have long-term survival.[10-12] In the perspective of insurability, it may be used to address the prognosis at the time of insurance inception. However, conditional survival does not discriminate individuals who are free-from-cancer from those who have already relapsed and are not amenable to insurance subscription.

Breast cancer stands as a paradigm where incidence is high, tumor heterogeneity is important, and curability is such that many patients will become long-term survivors. This study was performed to develop methods for inclusive, accurate, and justifiable estimation of risk and rating for breast cancer survivors. In this study, we investigated parameters that can predict the risk of death in people who have history of cancer but still disease-free at any time after the diagnosis of breast cancer using several machine learning algorithms. The use of large medical data sets often requires focusing to identify methods that can be used to analyze high-dimensional, heterogeneous, survival data with data sets that contain missing data or information such as comorbidities. While this may be regarded

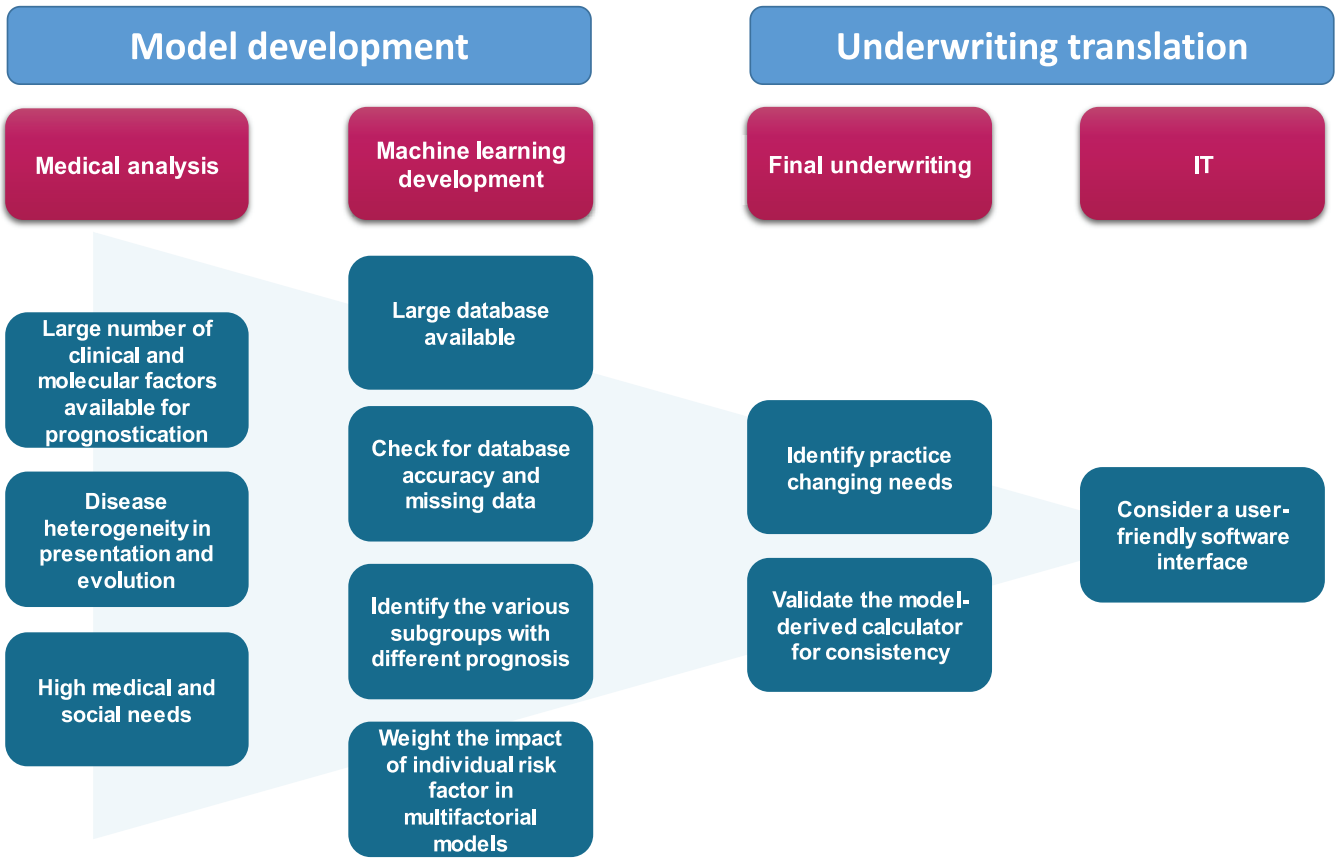

Figure 1. *Stepwise medical knowledge and actuary machine learning processes developed for underwriting translation.*

as a limitation, some authors are suggesting that including "missing information" is in fact heavily relevant to derive information that can be readily applicable to large and unselected patient populations. Previously published data and our own experience have suggested that machine learning can provide more accurate alternatives to traditional methods for survival analysis in the presence of high-dimensional data. For instance, the Cox proportional hazards model was previously described as poorly designed to handle high-dimensional data because the large number of features in the analyzed data causes the model to overfit the data, thus reducing the accuracy of the results for unseen data. Based on previously published experience, we decided to primarily focus our analyses using classification rather than more common survival time algorithms. This work was initiated based on models initially developed at SCOR by Thibault Antoine and Razvan Ionescu. The herein described method is only a part of a comprehensive process for product development, constantly challenging medical knowledge in oncology for frequent diseases with heterogeneous presentation and difficult prognostication (such as breast cancer) but with large available data sets to ensure the development of machine learning tools with the final aim of developing underwriting tools that can be made available through user friendly interfaces (Figure 1). From this investigation, we subsequently developed a breast cancer calculator that can now be used by insurance underwriters to estimate breast cancer survivor risks and to adjust individual ratings with history of breast cancer.

This paper was not aimed to describe and select a specific model to be readily used for reader but was rather aimed to provide

insight on methodologies that can be used to reach such a model. Thereby, the presented logistic regression model was taken as an example to support the stepwise development of this methodology for various medical applications, understanding that individual working in the field may be developing other modelizations.

METHODS

Database Selection

The SEER (Surveillance, Epidemiology and End Result) database has been used (https://seer.cancer.gov/data/). This database has been collecting data since 1973 and stands as the world's largest database specialized in cancer and recognized by the entire scientific community.[13] More than 400,000 observations are added every year. In terms of breast cancer, SEER has recorded more than 1.6 million observations of breast cancer since its creation and each observation is represented by 133 variables. Biometric variables (age at diagnosis, ethnicity, marital status, etc), medical variables (tumor size, tumor stage, etc) or therapeutic variables (surgery, chemotherapy, etc) are defined.

The reasons that led us to use this database included its size (the largest available cancer database in the world representing 30% of the American cancer population in US), its representativity (this database is representative of the American population in terms of socio-professional criteria), its reliability (partnerships with several laboratories and government agencies ensure the reliability and accuracy of the data), and its popularity (this database is used by physicians, researchers, scientists, and statisticians to conduct their work and publish figures on cancer incidence, prevalence, and mortality).

Survival of *in-situ* (non-invasive) breast carcinomas were not considered in this study and metastatic breast cancer risk estimation were not calculated because, while anyone can apply for insurance, only those in remission (ie, with no metastatic relapse) are generally considered candidates for insurance.

The selection was arbitrarily made to look primarily at death occurring within 10-15 years of follow up as this duration is frequently used in insurance as a reference for cancer risk evaluation.

Constraints to Apply to Select Variables

Variables were selected based on the following constraints: (1) *Business constraints:* Variables must be easily available to insurance companies. This means that the variable must be included in the insured's medical file at the time of application. For management issues, the number of variables that the underwriter must fill in the calculator must be kept reasonably limited. (2) *Commercial constraints*: The consistency of prices is a fundamental aspect for underwriters. Rates must be consistent with the medical knowledge, meaning that the extra premium must be lower for an individual with a small tumor size than one with a larger tumor. (3) *Medical constraints*: The selected variables must also be considered as prognostic variables that are clearly identified by oncologists and recognized as such in the medical literature. (4) *Statistical constraints*: The chosen variables must have a high "Feature Importance." That means that it must have a significant impact on prediction.

The prognostic value of selected variables was further evaluated according to the Kaplan Meier estimate of survival.

The outcome of metastatic breast cancers was used in the method to evaluate the probably of death following metastatic relapses but were not used for rating, primarily because the estimated life duration in metastatic breast cancer was too short to make any relevant insurance offer in those patients.

Risk of Relapse Model

The following considerations were used to build up a model (Figure 2):

1. A person in cancer remission (ie, after treatment) may relapse into this disease.

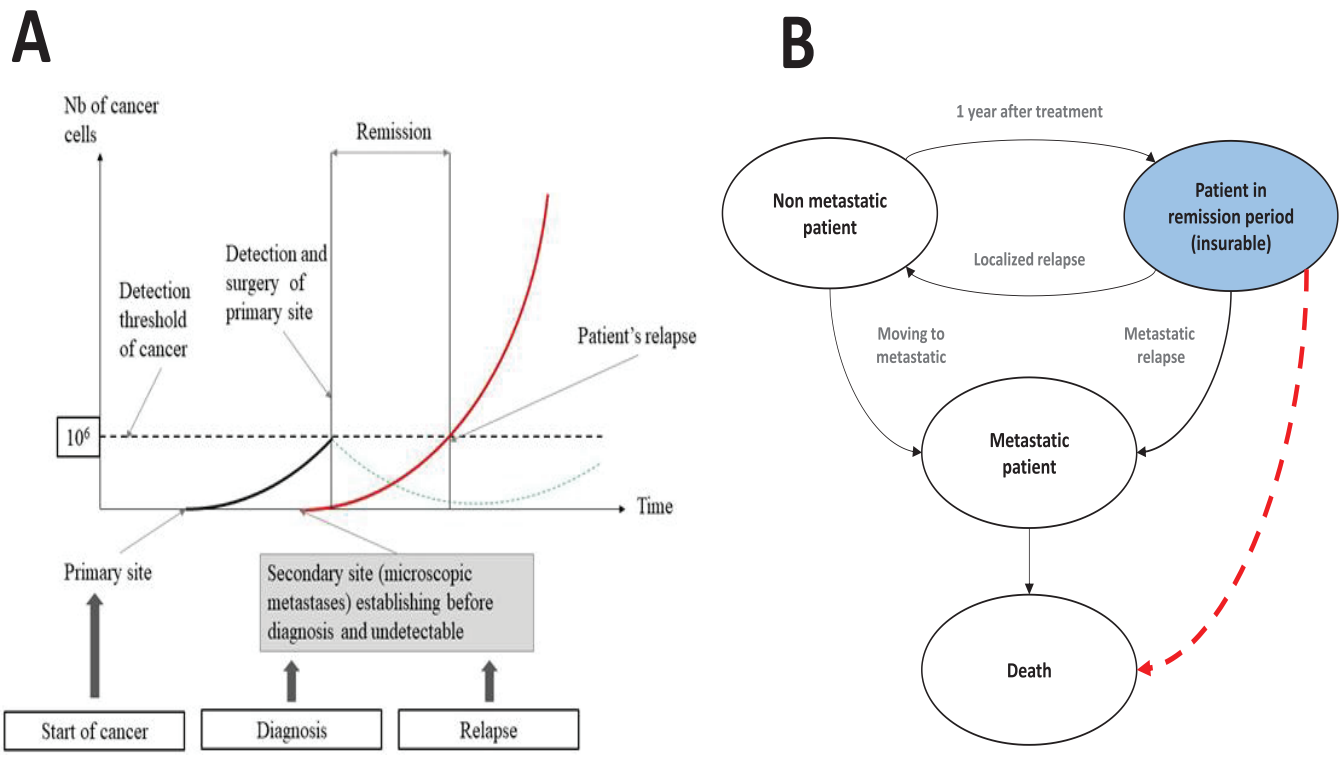

Figure 2. *Natural history of breast cancer evolution (A) and development of a model where patients in remission (long-term survivors) after treatment may consider applying for insurance (B).*

2. A person with breast cancer (or in remission) can only die of breast cancer if he or she is metastatic (or has a metastatic relapse).
3. The longer the remission period, the lower the risk of relapse.

Importantly, the "metastatic relapse" is not updated in the database. Thus, in SEER database, a non-metastatic person who dies from specific breast cancer death is considered to have necessarily relapsed into metastasis during the observing period.

It is also important to note that only insured people in remission will be included in this analysis as only individual in remission will be able to apply for insurance. Thus, the statistical tool must consider the fact that insureds are exposed to the risk of relapse. In addition, it is considered that the longer the remission period, the lower the risk of extra mortality.

We estimated that those who died specifically from breast cancer using the SEER database, must have experienced a relapse at some point before death. This point was determined using the survival estimate of those with metastatic cancer at diagnosis. At the estimated point of relapse, these individuals were removed from analyses in relevant future time points. We estimated the risk of death of patients with metastatic breast cancer according to the number of metastatic lymph node whenever available, considering that lymph nodes are the most important prognostic parameters.

Models selected for analyses were aimed to consider medical findings and the limitations of the database, a model which considers this risk of relapse and the conditional survival (the probability of being alive knowing y years without having relapsed).

Thus, the calculation of the extra premium associated to each risk profile was performed in 5 steps (Figure 3):

Step 1: To determine the probability q_x of" specific breast cancer death" for non-

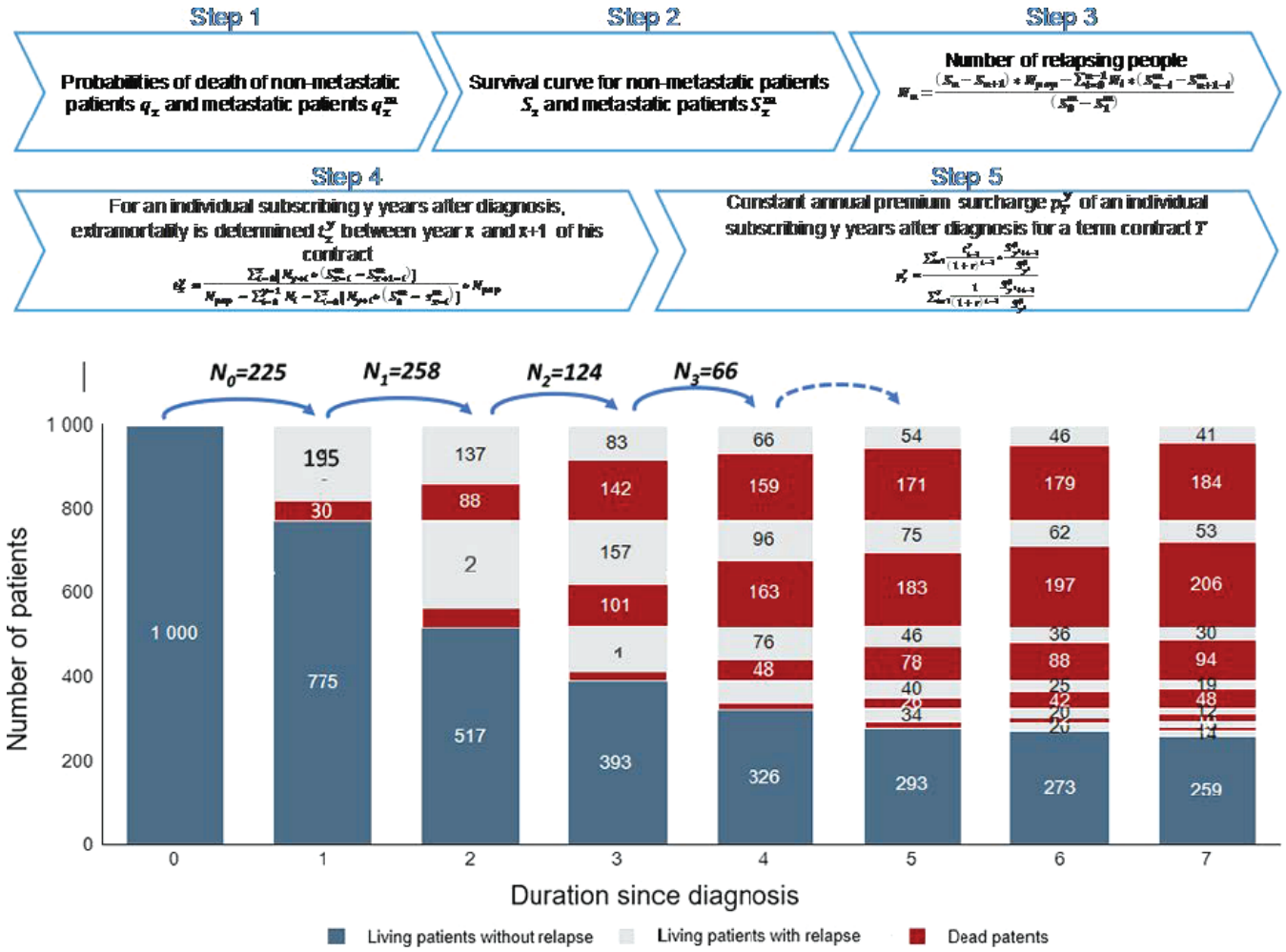

Figure 3. *Stepwise development of a breast cancer model of conditional survival for cancer-free patients applying for insurance.*

metastatic observations and the probability q_x^m of dying from" all causes" for metastatic observations, x years after the diagnosis. With x = {1, ..., 15} the time since diagnosis.

Step 2: To plot the survival curves S_x and S_x^m, respectively the survival curve for non metastatic and the survival curve for metastatic x years after the diagnosis, using the predictions from step 1 with $S_{x+1} = S_x \times (1 - q_{x+1})$ and $S_0 = 100\%$.

Step 3: To determine the number of relapses N_n that occurred each year n of remission with:

$$N_n = \frac{(S_n - S_{n+1}) * N_{pop} - \sum_{i=0}^{n-1} N_i * \left(S_{n-i}^m - S_{n+1-i}^m\right)}{\left(S_0^m - S_1^m\right)}$$

Step 4: To determine the extra mortality t_x^y(which is equivalent to the extra premium) y years after the diagnosis and x years after the inception with:

$$t_x^y = \frac{\sum_{i=0}^{x} [N_{y+i} * \left(S_{x-i}^m - S_{x+1-i}^m\right)]}{N_{pop} - \sum_{i=0}^{y-1} N_i - \sum_{i=0}^{x} [N_{y+i} * \left(S_0^m - s_{x-i}^m\right)]} * N_{pop}$$

This value represents the excess mortality between the year x and x+1 of the contract of an individual subscribing after y years of remission. However, these extra mortalities are not the final extra premium that the insured should pay because they are not constant.

Step 5: To compute the constant extra premium p_T^y of an individual subscribing y years

after diagnosis for a T duration of contract. An "actualized" and "probabilized" average of these extra mortalities can be calculated to define the constant annual premium surcharge that the insured will have to pay over the period T of the contract. So, this constant extra premium is defined as follow:

$$p_T^y = \frac{\sum_{i=1}^{T} \frac{t_{i-1}^y}{(1+r)^{i-1}} * \frac{S_{y'+i-1}^g}{S_{y'}^g}}{\sum_{i=1}^{T} \frac{1}{(1+r)^{i-1}} \frac{S_{y'+i-1}^g}{S_{y'}^g}}$$

Presentation of Algorithms and Metrics

The first objective was to determine the probability of death of an individual with breast cancer using various methods of machine learning.[14,15] This probability is thought to depend on the selected prognostic variables. The following machine learning processes were applied: Logistic Regression, GA2M, Random Forest, and XGBoost.

- Regression Logistic: Parametric model of the GLM family, it approaches each variable x_j by a vector β_j. Then using its link function logit, the probability of death of everyone is deduced.
- Random Forest: This algorithm is a model based on decision trees. However, instead of using a single decision tree (CART), the Random Forest creates several trees with re-sampling (Bagging), to limit overfitting. Then it aggregates the predictions of all the predictions to create one.
- XGBoost: Also based on the principle of decision trees, it uses the Boosting method to make these predictions. This means that it aggregates different predictors (or trees) sequentially to correct its predictions.
- GA2M: It is also a parametric model, of the GAM family. It approaches each variable with a parametric function. But the particularity of this model is that it considers the interactions between each pair of variables, in its link function logit.

Table 1. Patient Characteristics

Total number of patients in the SEER data set	**942,188**
Total number of patients in the selected population with age range 18-80 years old	837,779
Gender	
Female	99.3%
Male	0.7%
Pathological characteristics	
Ductal	75.5%
Lobular	18.5%
Other	6.0%
TNM stages	
T stages	
T1a	8.8%
T1b	17.3%
T1c	34.2%
T2	27.2%
T3	4.8%
T4	2.7%
Unknown T	5.1%
N stages	
N0	65.7%
N1	22.7%
N2	6.1%
N3	4.5%
Unknown N	1.0%
M stages	
M0	94.9%
M1 or MX	5.1%

SEER: Surveillance, Epidemiology and End Result, T, N, and M refer to Tumor, Node, and metastasis used in the TNM classification.

The performances of these algorithms have been compared with metrics such as Log Loss (evaluating the accuracy of predicted probabilities), AUC (evaluating the accuracy of classifications), SMR (evaluating the calibration of death probabilities), loss Ratio (evaluating the profitability of pricing), and Consistency tests. GG-plot, GGsurv-plot and SurvMiner packages were used as analytical tools. For building models, we use the Python programs. As we have been looking at various models, we didn't consider specifically the end odd ratios because it was not applicable in all cases.

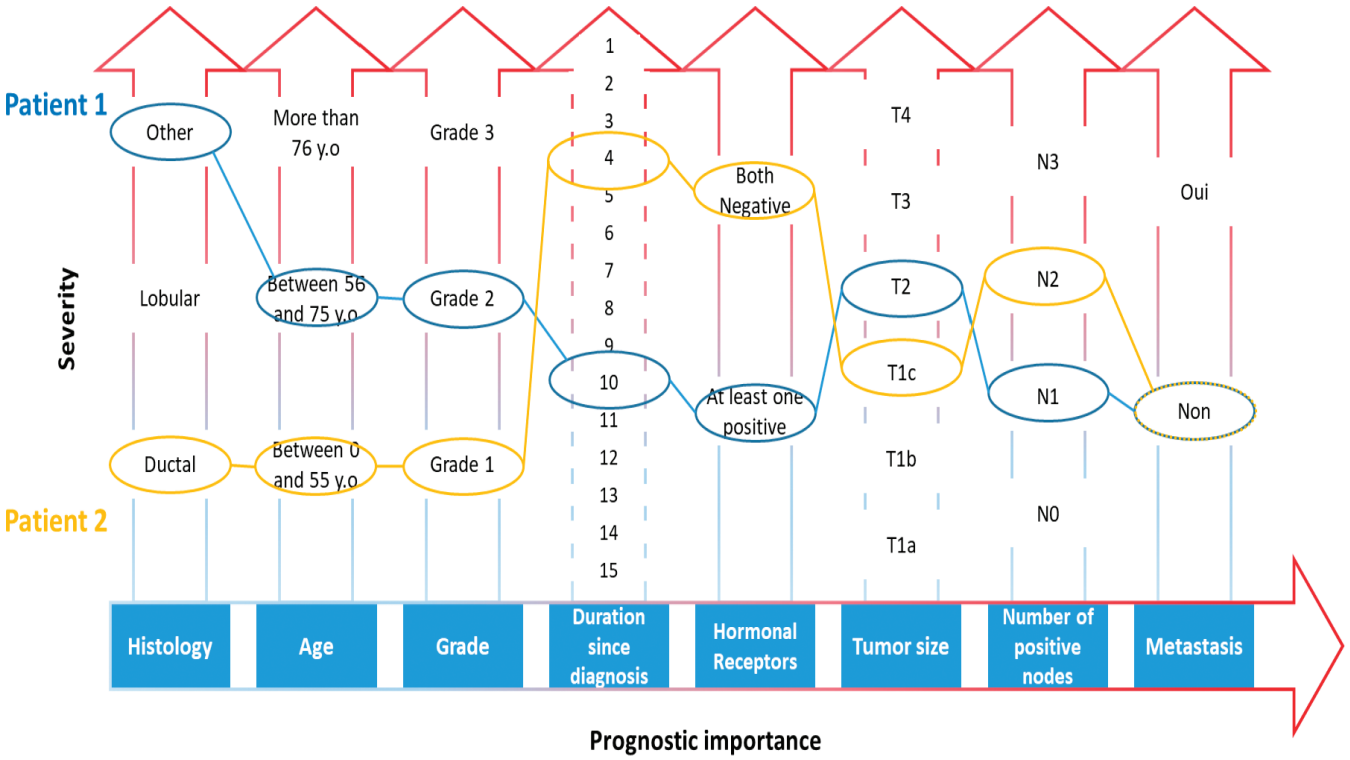

Figure 4. *Typical examples that encompass multiple parameters involved in prognostication of breast cancer risk.*

RESULTS

Selection of Variables Associated with a Risk of Relapse and Death

The population of patients entered into this study is described in the Table 1. Age is indeed a great predictor of mortality independently of breast cancer. Moreover, breast cancer occurs more frequently at an age where standard mortality occurs. In this work, results were adjusted by age, allowing the estimation of the over risk associated with breast cancer. Among multiple parameters selected to predict the survival outcome (Figure 4), The following 7 clinical parameters that can be identified from the pathological report after surgery were selected: the pathological subtype, pTNM staging (T: tumor size, N; number of nodes; M presence or absence of metastases), Scarff-Bloom-Richardson grading, the expression of estrogen and progesterone hormone receptors and the duration since diagnosis.

Those parameters fulfilled constraints that applied to variables with low rate of missing data in the SEER database. Therefore, those parameters may be considered as the most representative whatever the model (Figure 5). Importantly, the duration between the diagnosis and the onset of insurance inception as well as the age at subscription were also important to consider (the longer the interval duration the lower the risk of relapse) and was included on the model.

Selection Among Machine Learning Algorithms

The robustness of the Logistic Regression, GA2M, Random Forest, and XGBoost models was assessed by comparing them with the Kaplan Meier estimator. All the predictions made using the 4 models were close to those of the Kaplan-Meier estimator and survival curves (Figure 6).

The Area Under the Curve of the Receiver Operating Characteristic (ROC) is between 0.78 and 0.80 whatever the model (Figure 7).

Moreover, predictions from XGBoost, Logistic Regression and GA2M were confused

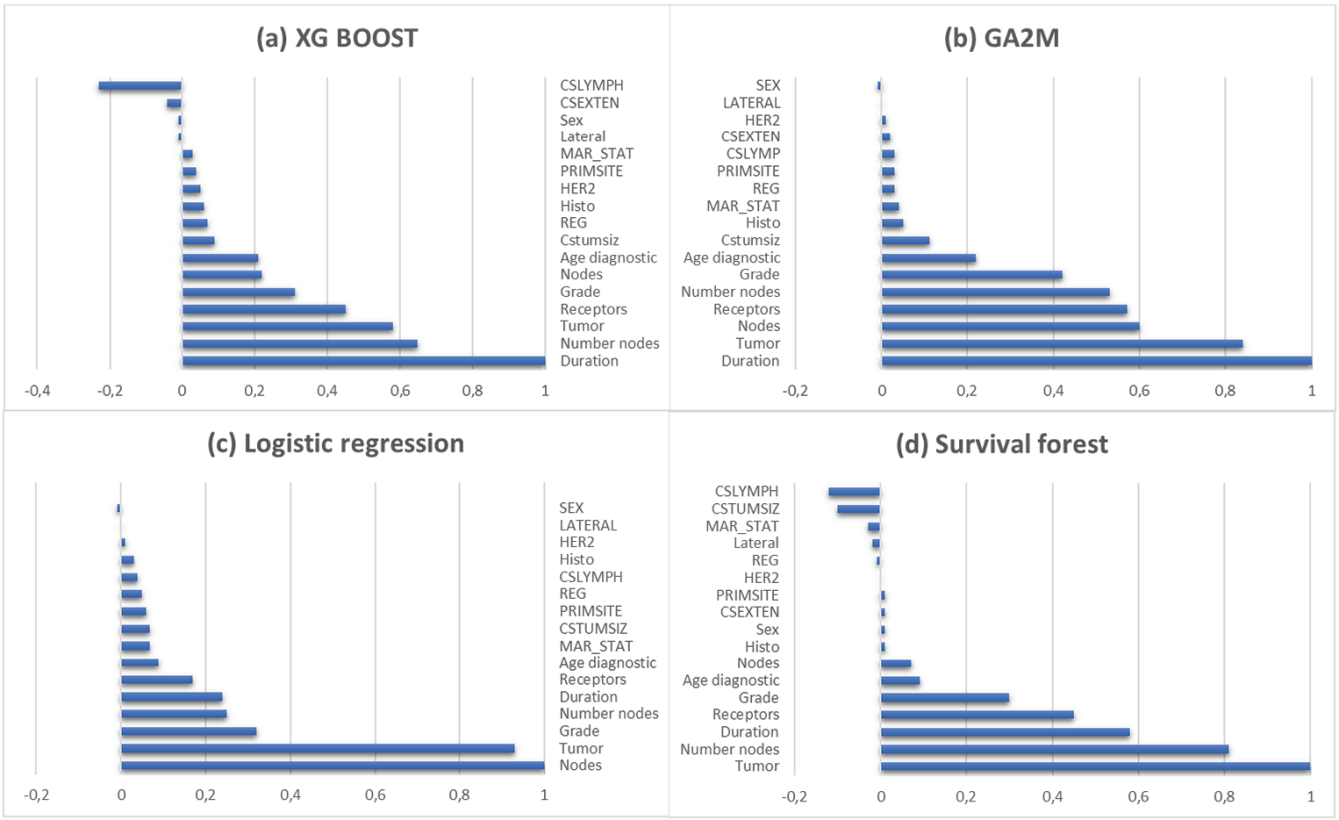

Figure 5. *Relative rankings of prognostic parameters in various models of machine learning.*

for non-metastatic observation, shows that they predicted individual outcome in the same way. For metastatic observations, predictions are near to the Kaplan-Meier estimator, but have not confused each other. While all models had identical performance in terms of statistical metrics (AUC, Log Loss, and SMR), the logistic regression was the one and only model that respects all business constraints and was intelligible for medical and underwriting users.

Setting Up a Breast Cancer Calculator

Based on the above-mentioned model, we have set a breast model calculator using easy to recover information from patients with a history of breast cancer. The model estimates from the selected variable the 15-year conditional disease-free survival rate at any time from diagnosis and can, therefore, be used for individuals with a history of breast cancer, the individual risk of death, and define individual risk-adjusted pricing.

Specific Medical Situations

Two medical situations were shown to require particular attention to use the breast cancer calculator. One is the risk estimation of synchronous multifocal lesions occurring in the same breast. In this situation, the calculation shall consider the risk of each separate breast lesion, considering the overall risk estimation, the lesion of poorest prognosis. The other peculiar situation is the use of neoadjuvant chemotherapy for locally advanced/poor prognosis breast cancers that is expected to downsize the breast tumor (ypTNM), and this downstaging is likely to lead to underestimation of the risk. In this later situation, parameters to consider shall not be recovered from the surgical pathological report but from the clinical informa-

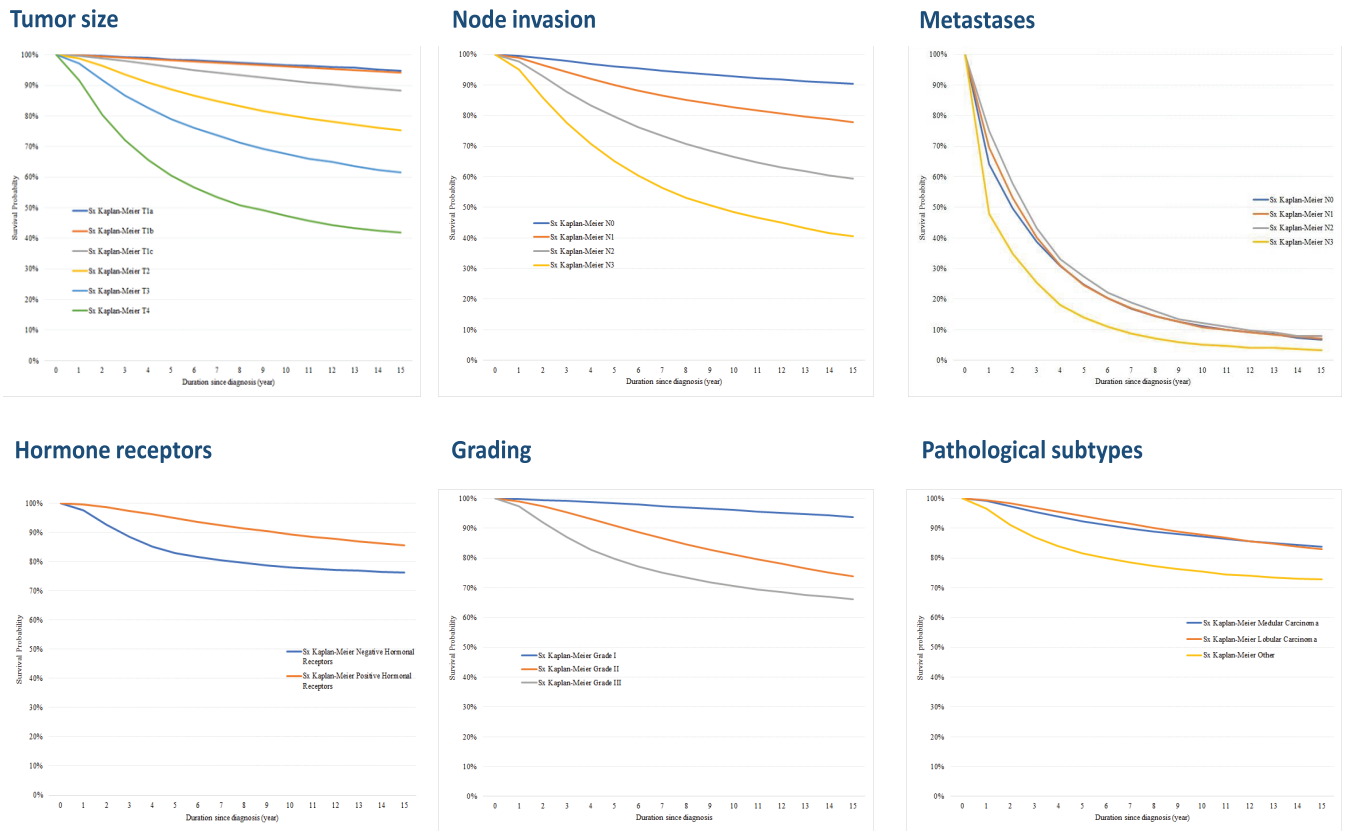

Figure 6. *15-year overall survival of breast cancer patients according to the main prognostic factors based on the TNM classification, the presence of hormone receptors, the Scaff Bloom Richardson grading, and pathological subtypes.*

tion cTNM staging and the biopsy performed prior to any chemotherapy.

DISCUSSION

This project was aimed at providing a global method to assist worldwide underwriters in accurately addressing the risk-derived rating for individuals with a history of breast cancer using data-driven evidence based on machine-learning algorithms. Data were obtained from the SEER database that accounts for the largest, regularly updated, dataset of patients with cancer. As statistical analyses of breast cancer had to be consistent with the most advanced medical knowledge, we ensured that the process used in this project, as well as the risk evaluation, fulfills criteria that could be acceptable by multiple parties by validating each step in our team that included at least one of the following experts: a statistician, an actuary, an underwriter, a programmer, and a medical oncologist.

The process developed in this project was shown to fit well for breast cancer that stands as a high incidence malignancy. Importantly, large breast cancer databases are available for statistics. Finally, insurance needs derive from the high heterogeneity of breast cancer for which the outcome may vary depending off several prognostic factors. As a word of caution, such machine learning development may not fit for other cancer types for which incidence is lower, a large database is not available, and/or for which prognostic parameters are limited. In this later situation, risk estimation based on survey of published medical data and medical intelligence may be sufficient to regularly revised ratings (Figure 8).

In this study, we redefined parameters predicting survival in individuals with a history of breast cancer at any time point from di-

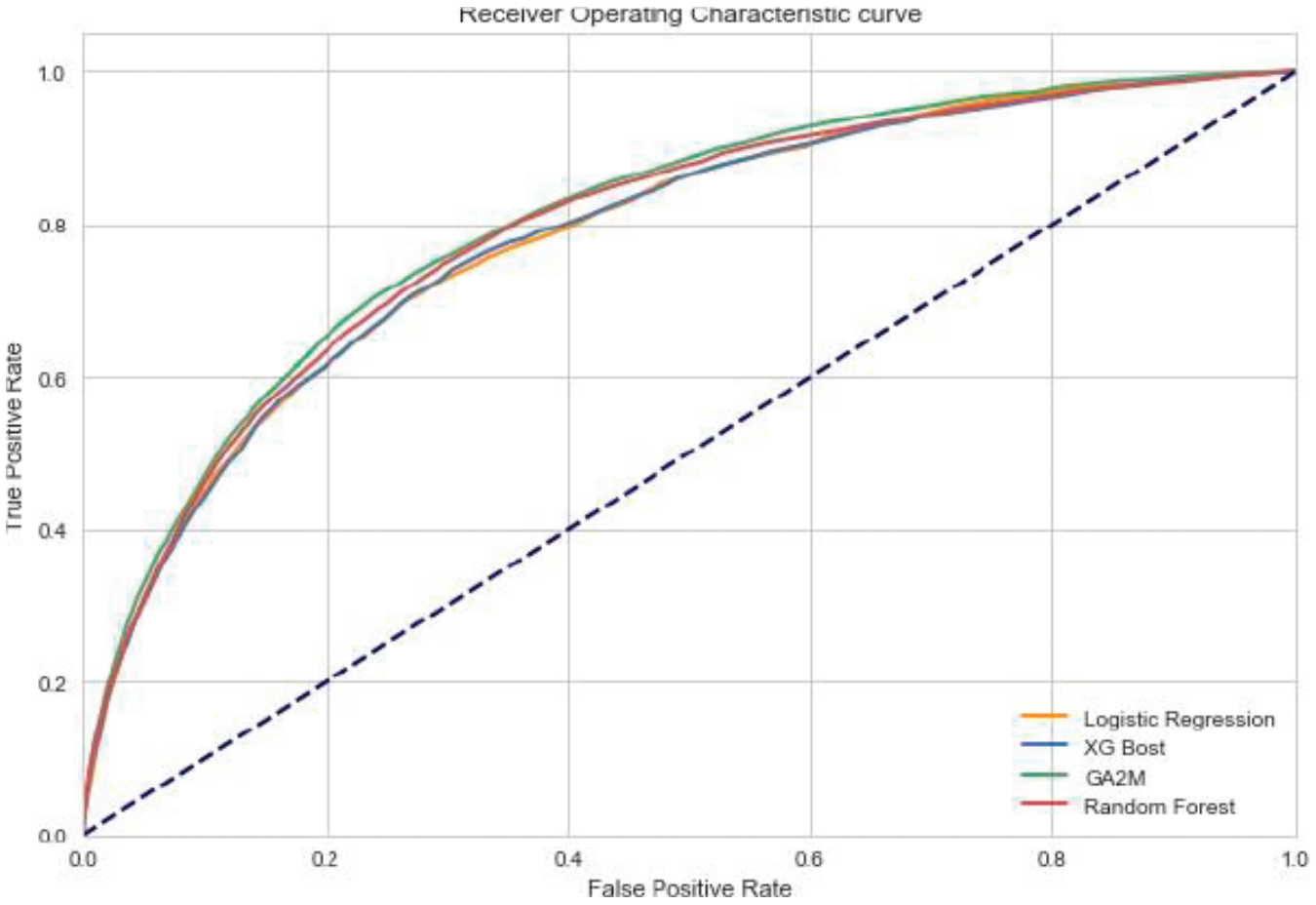

Figure 7. *Performances of four machine learning models (Logistic regression, XG Boost, GA2M and Random forest).*

agnosis providing they are still disease-free at the time of insurance inception. As expected, the duration since diagnosis was associated with survival estimates, the longer the disease-free survival duration, the better the decreased risk of further death. Furthermore, the age at the contract signature and the duration of the contract inception were also important factors to estimate the insurability.

Although variables primarily used for diagnosis were evaluated, not all medical parameters used at diagnosis by treating physicians for prognostication were relevant in our model, years after diagnosis. Relevant medical parameters already evaluable at the time of diagnosis and still relevant for our calculator were the histological subtype, pTNM classification, SBR grading, and hormone receptor expression. In our model, the overexpression of HER2-receptors, that is acknowledged as a factor of poor prognostic at diagnosis, was not associated with subsequent poor outcomes, which was eventually due to the systematic use of adjuvant antibody-based therapy with trastuzumab (Herceptin ™) targeting HER2 3+ positive cancer cells for 1 year.

Interestingly, prognostic estimates using disease-free conditional survival at the time of insurance inception was shown to be consistently better than prognosis at the time of diagnosis. This reduction of the risk of death overtime has been described elsewhere and is mainly related to the occurrence of harvesting death of poor prognostic patients leading to an apparent improvement of the estimated survival in still disease-free patients. In our study, this has led to an improved estimated of risk and rating overtime, allowing better rating offers for long-term disease-free breast cancer survivors. As eluded to above, the du-

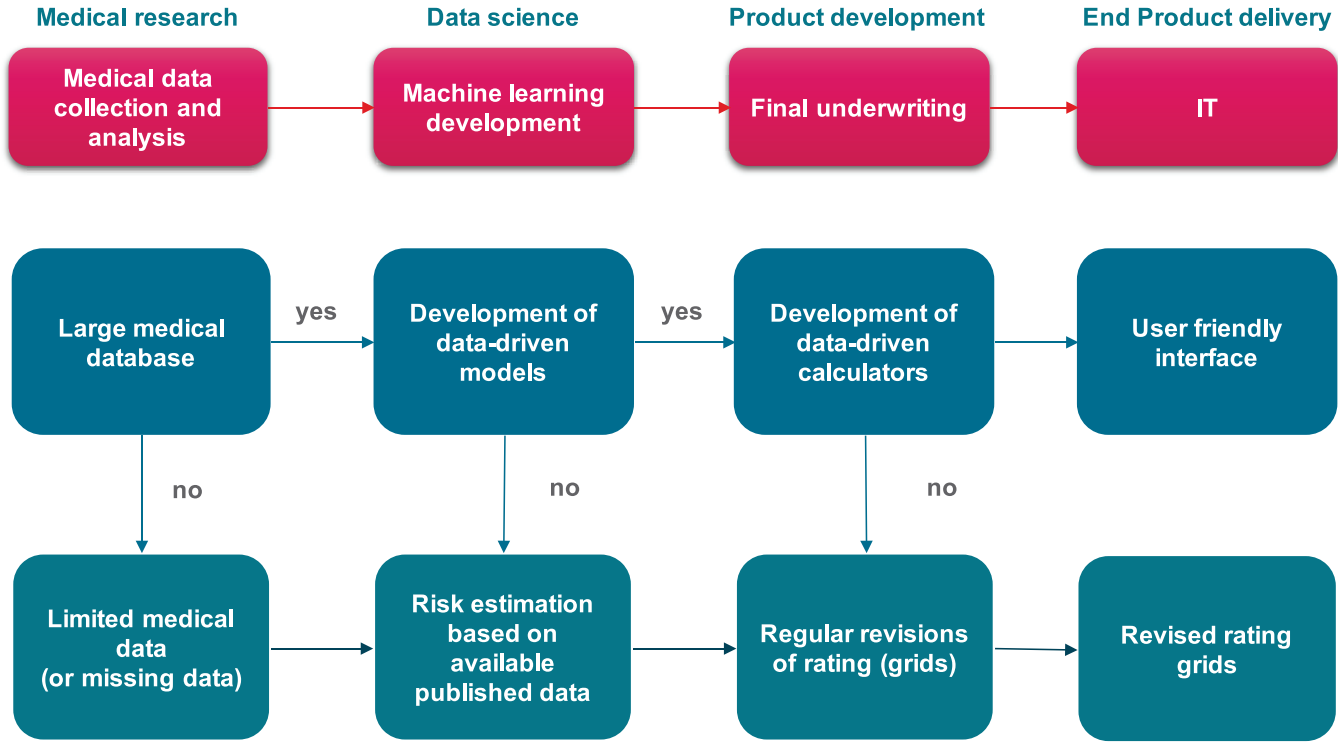

Figure 8. *Development of novel underwriting products either based on machine learning and calculators or more regular processes based on medical intelligence.*

ration since diagnosis appears to be an important parameter that has been subject to further political debates in many countries such as France where *the right to be forgotten* after 10 years of disease-free survival has been set by law for breast cancer survivors. Independently, our model was shown to better predict survival of breast cancer survivors allowing a more accurate estimation of risk that eventually will lead to better ratings.

Our method was set to propose individualized risk estimation to derive justifiable pricing based on 7 *easy-to-recover* characteristics, which are: age, duration since diagnosis, size of tumor, number of lymph nodes affected, grade of tumor, hormone receptors, and histology. These variables were selected with respect to medical constraints (medical relevance of these variables), underwriting constraints (accessibility of these variables at the time of underwriting), and statistical constraints (important variable for all models). Thus, our method appears consistent with the expected heterogeneity of breast cancer risk with more than 20,000 combinations of variables, making it possible to derive adjusted rating for each specific case.

The main challenge was the choice of the prediction model to make the pricing. While all models had similar performance in terms of statistical metrics (AUC, Log Loss, and SMR), the logistic regression was the only model that respects all business constraints and was intelligible for users. In fact, the coherence with pricing will be probably the most stringent constraint to be usable by underwriters.

CONCLUSION

The development of more inclusive underwriting requires a comprehensive knowledge of the heterogeneity of breast cancer. Recent advances in Machine Learning and the availability of rich databases specialized in breast cancer were shown to allow significant improvement in data-driven risk predictions. In this study, we were able to

develop a method for an accurate estimation of individual risks, which further lead to the global development of an insurance specific and underwriter-friendly calculator with justifiable pricings for people in remission from breast cancer.

REFERENCES

1 Fitzmaurice C, Abate D, Abbasi N, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Lif.e Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol.* 2019;5:1749-1768.

2 Nardin S, Mora E, Varughese FM, et al. Breast Cancer Survivorship, Quality of Life, and Late Toxicities. *Front Oncol.* 2020;10:864.

3 Bodai BI, Tuso P. Breast Cancer Survivorship: A Comprehensive Review of Long-Term Medical Issues and Lifestyle Recommendations. *Perm J.* 2015;19:48-79.

4 Thong MSY, Doege D, Weißer L, et al. Health and life insurance-related problems in very long-term cancer survivors in Germany: a population-based study. *J Cancer Res Clin Oncol.* 2022;148: 155-162.

5 Vromans RD, van Eenbergen MC, Geleijnse G, Pauws S, van de Poll-Franse LV, Krahmer EJ. Exploring Cancer Survivor Needs and Preferences for Communicating Personalized Cancer Statistics From Registry Data: Qualitative Multimethod Study. *JMIR Cancer.* 2021;7:e25659

6 Eloranta S, Smedby KE, Dickman PW, Andersson TM. Cancer survival statistics for patients and healthcare professionals - a tutorial of real-world data analysis. *J Intern Med.* 2021;289:12-28.

7 Yang H, Pawitan Y, He W, et al. Disease trajectories and mortality among women diagnosed with breast cancer. *Breast Cancer Res*. 2019;21:1-8.

8 Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol.* 2019;19:64.

9 Pruessmann J, Pursche T, Hammersen F, Katalinic A, Fischer D, Waldmann A. Conditional Disease-Free and Overall Survival of 1,858 Young Women with Non-Metastatic Breast Cancer and with Participation in a Post-Therapeutic Rehab Programme according to Clinical Subtypes. *Breast Care (Basel).* 2021;16:163-172.

10 Iraji Z, Asghari Jafarabadi M, Jafari-Koshki T, Dolatkhah R. A Conditional Probability Model to Predict the Mortality in Patients with Breast Cancer: A Bayesian Network Analysis. *Am J Med Sci.* 2020;360:575-580.

11 Ai B, Wang X, Kong X, Wang Z, Fang Y, Wang J. Conditional Survival of female patients with operable invasive Breast Cancer in US: A population-based study. *J Cancer.* 2020;11:5782-5791.

12 Leone JP, Vallejo CT, Hassett MJ, et al. Factors associated with late risks of breast cancer-specific mortality in the SEER registry. *Breast Cancer Res Treat.* 2021;189:203-212.

13 Zhong M, He X, Lei K. Survival of Patients with First and Metachronous Second Primary Breast Cancer or Lung Cancer Malignancy: Comparisons Using the SEER Database. *Adv Ther.* 2020;37:2236-2245. https://seer.cancer.gov/data/

14 Boeri C, Chiappa C, Galli F, et al. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Med.* 2020;9:3234-3243.

15 Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast Cancer Prognosis Using a Machine Learning Approach. *Cancers (Basel).* 2019;11:1-9.