SCOR
The Art & Science of Risk

A NEW CALCULATOR

aimed at providing fair insurance cover to people with a breast cancer history

MOLL Antoine PLISSON Manuel RAYMOND Eric SMITH Eimear

Authors:

Antoine Moll
Senior Data Analyst

Manuel Plisson, PhD
Head of Biometric Modelling & Inclusive Underwriting

Dr. Eric Raymond
Professor of Medical Oncology & Chair of the Medical Oncology Department at Paris Saint-Joseph Hospital

Eimear Smith
Senior Technical Underwriting Consultant

Introduction and Context

Medical and epidemiological analysis at international level shows that breast cancer represents about **25% of all cancers in women** and affects more than one in eight women in the course of their lives, making it the most common cancer in the world. With the improvement of early diagnosis and screening methods, often encouraged and subsidised by cancer associations and health authorities, **breast cancer tends to be diagnosed earlier, at stages associated with very good prognosis.** Hence, 90% of people diagnosed with breast cancer are still alive 5 years later. Recent medical advances in treatment and screening are leading to more and more remissions from breast cancer (long-term survivors) and it is our strong belief that people with a history of breast cancer are now often eligible to be insured. This echoes already existing Underwriting programs and Regulations in different countries, that have allowed for a greater inclusion of life insurance applicants with a history of breast cancer.

Defining the individual prognosis of a person with breast cancer to assess the risk of relapse and death is medically associated with several clinical and biological prognosis parameters that are used by doctors at diagnosis and can be found in medical charts. The development of a more inclusive underwriting program for people with a breast cancer history **requires a very detailed knowledge of these prognostic parameters.** Moreover, with the recent advances in Machine Learning and the availability of specialised rich data, the quality of predictions should improve significantly.

This is the reason why we have developed **VITAE Breast Cancer. Using state of the art technologies, actuarial techniques and medical knowledge, this calculator aims to assess the risk of people diagnosed with breast cancer based on a wide range of parameters that are good predictors of recurrence and death. This allows for a more accurate estimation of individual risk to better adjust underwriting.** A special focus has been put on making that tool as user-friendly as possible for Underwriters.

What Vitae Breast Cancer allows us to do

A SINGLE TOOL TO ASSESS BREAST CANCER CASES DESPITE THEIR HETEROGENEITY

Calculators such as Vitae Breast Cancer simplify underwriters' work as they allow us to gather all the information necessary to compute the risk on a user-friendly, single platform. Indeed, some key information must be considered when going through the application of a person in breast cancer remission (see across).

Risk factors used for underwriting of people with a breast cancer history

- Type of histology;
- Current age;
- Grade;
- Hormonal receptors;
- Duration since diagnosis;
- Size of the tumour;
- Number of nodes.

The list above may appear short but is crucial in terms of importance in the evaluation of the vital prognosis (metastasis status being the riskier). As each variable has several factors, around 20,000 combinations of these variables exist. This makes breast cancer one of the tumours for which a "multifactorial" risk evaluation justifies the need for a calculator. To facilitate risk assessment, a specially designed user interface guides underwriters through all the steps of the risk assessment, providing explanations for each risk factors.

A CLEARLY DEFINED PERIMETER OF ACCEPTED TYPES OF BREAST CANCER

Vitae Breast Cancer is built to assess the risk of most breast cancer cases[1]. However, the calculator can't propose cover for the riskiest cases as they are **not insurable**. Some other impairments[2] should also be assessed with specific grids. **For all other cases, Vitae Breast Cancer empowers underwriters to assess risks and propose an outcome.**

Vitae Breast Cancer is based on data and validated with medical expertise

THE SEER DATABASE PROVIDES A WEALTH OF INFORMATION ON BREAST CANCER

The Vitae Breast Cancer algorithm was built using the SEER (Surveillance, Epidemiology and End Result) database. This American database has been collecting data since 1973. Today, SEER is the world's largest database specialised in cancer and recognised by the global medical and scientific community. More than 400,000 observations are added every year. In terms of breast cancer, SEER has recorded more than 1.6 million observations since its creation and each observation is represented by 133 variables. Biometric variables (age at diagnosis, ethnicity, etc.), medical variables (tumour size, tumour stage, etc.) or therapeutic variables (surgery, chemotherapy, etc.) are provided.

The following reasons have led us to select this database:

- **Reliability.** SEER data is built through partnerships with several laboratories and government agencies, ensuring the reliability and accuracy of the information gathered.
- **Scientific validation.** This database is widely used by researchers and statisticians to conduct their work and publish figures on cancer incidence, prevalence and mortality.
- **Size.** It is the largest cancer database in the world, ensuring statistical power of results.
- **Representativity.** This database is representative of the American population in terms of socio-professional criteria.

ADDITIONAL FEATURES AND REQUIREMENTS FOR VITAE BREAST CANCER

To provide the most adapted tool for underwriters, the Vitae Breast Cancer calculator was built to comply with several constraints:

1. Business

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. The number of variables that the underwriter is required to fill in the calculator must be reasonable.

2. Commercial

The consistency of prices is a fundamental aspect for underwriters to justify loadings. Rates must be consistent with medical discourse. For example, if the medical literature indicates that an individual is more likely to survive with a 5 mm tumour than with a 20 mm tumour, then the fee should follow the same logic. In other words, the extra premium must be lower for an individual with a 5 mm tumour than if they had a 20 mm tumour.

3. Medical

The selected variables must also be considered as prognostic variables that are clearly identified by physicians and recognised as such by the medical literature.

4. Statistical

The chosen variables must have a high feature importance. In other words, elements included in the model must have a significant impact on predictions.

Finding the best method to model Breast Cancer mortality risk

ESTIMATING THE PROBABILITY OF DEATH

The first objective is to determine the probability of death of an individual with breast cancer. Econometrics and Machine Learning techniques were compared to find the more accurate method:

- **Logistic Regression:** Parametric model of the GLM family, it approaches each category of each variable with an estimate. Then using its logit link function, the probability of death of each individual is deduced.
- **GA2M:** 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.
- **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.

Performance of these methods are compared through model metrics (Log Loss, AUC and SMR), Loss Ratio actuarial metric and consistency tests of final prices:

- Model metrics are used to verify the accuracy of the model.
- The Loss Ratio estimates the profitability associated with final prices.
- Consistency tests aims to check prices are in line with medical literature, for example risk increases with size or number of nodes.

Though machine learning methods provide good predictions, the logistic regression was found to be the only that respects all consistency and business constraints.

TAKING INTO ACCOUNT THE RISK OF RELAPSE

The pricing model behind the Vitae Breast Cancer has been built by Thibault Antoine, Deputy Chief Knowledge Officer in charge of Data Solutions and Razvan Ionescu, Head of Biometric Risk Modelling at SCOR Global Life. It considers the risk of relapse. Taking into account metastasis and non-metastasis status permits the proposal of fairer prices in some cases. The model calculates the extra premium of an applicant with a history of Breast Cancer, i.e., the extra mortality of this person. In other words, it is the same as determining the probability that this person with a relapse dies specifically from this disease. Probabilities of death estimated above are used as the first step of the computation.

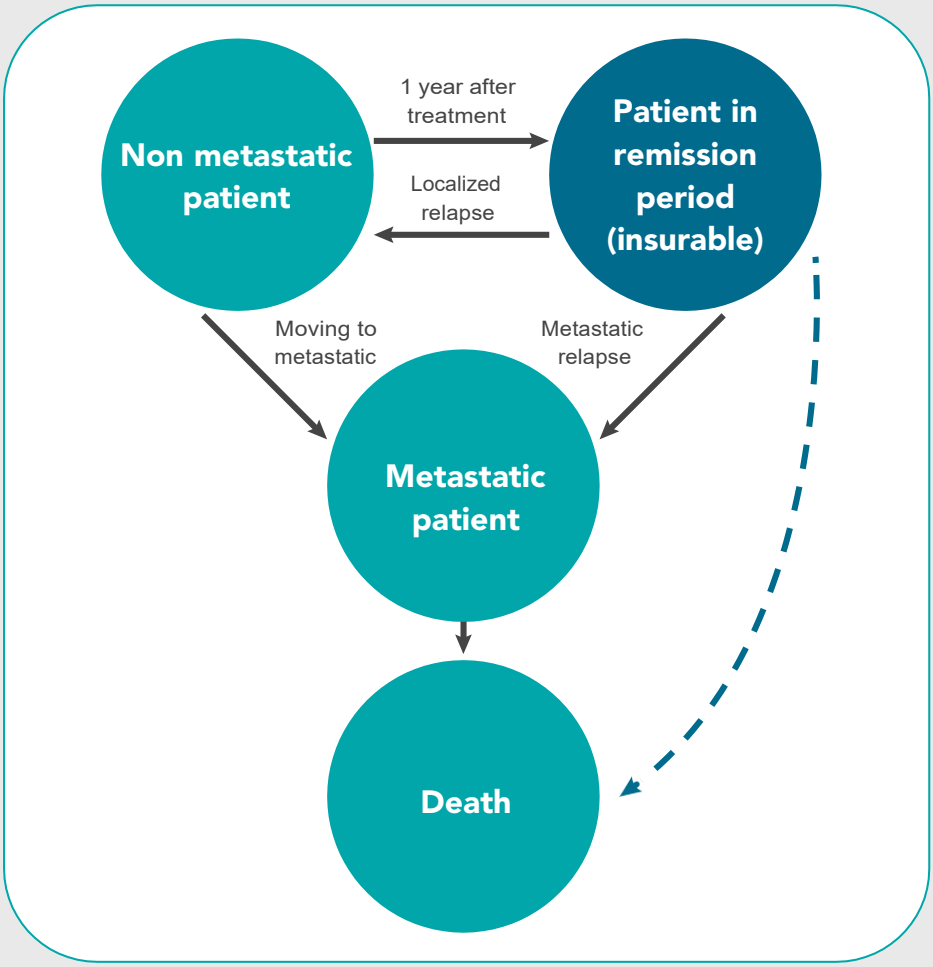
Extra premiums are expressed as per-millage (PM). The per-millage is used because the risk of breast cancer varies relatively little with age. For example, if an insured is subject to a surcharge of 4 PM per year for a guaranteed amount of $50,000, then each year of their contract they will have to pay a surcharge of $50,000 x 4/1000 = $200. This per-millage is the probability of this individual dying from the same disease.

Having a comprehensive database is just a first step. To provide the fairest pricing, the model is based on the following medical findings (see figure next page):

1. A person in cancer remission (i.e., after treatment) may relapse from this disease.

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.



However, "metastatic status" is not updated in SEER. Hence, a non-metastatic person who dies from specific breast cancer death is considered to have necessarily relapsed into metastasis during their observing period. More detail on risk of relapse is available in SCOR's Technical Paper of Vitae Breast Cancer.

In Brief

Our new calculator Vitae Breast Cancer provides underwriters a useful tool to assess risk. Combining actuarial, medical and underwriting expertise, it is in line with medical breakthrough and adapted to underwriters needs in terms of day-to-day work. Don't hesitate to contact us for further information and/or to test the tool.

1. Metastatic disease, inflammatory carcinoma, T4d, carcinomatosis mastitis or Pev III, locoregional recurrence or recurrence of a primary metachronous breast cancer

2. Erosive adenomatosis, papillary adenoma of nipple, spindle cell carcinoma of breast, breast sarcoma phyllodes – malignant, proliferative Aschoff's center, Radial scar, polycystic breast cancer, breast in situ carcinoma, Paget's disease, breast secreting carcinoma