A MORE INCLUSIVE CARDIOVASCULAR RISK CALCULATOR

based on latest medical research
and machine learning techniques

ABSTRACT
May 2021

Authors:

Antoine Moll
Senior Data Analyst

Manuel Plisson, PhD
Head of Biometric Modelling & Inclusive Underwriting

Dr. Pierre SABOURET, MD
Senior Research Fellow at the Pitié Salpétrière Hospital-Heart Institute and President of the Scientific Committee of the National College of Cardiologist of France.

Dr. Marinos FYSEKIDIS, MD
CMO Consultant in Diabetes, Endocrinology with subspecialty in Human Nutrition.

This document aims to briefly explain how we identified the most relevant risk factors to build Vitae Cardio, as well as the methodology we used to turn a rich data source into a powerful and robust underwriting solution.

Introduction and Context

Cardiovascular Diseases (CVDs) are the leading causes of death worldwide. An estimated 17.9 million people died from CVDs in 2019 according to the World Health Organization (WHO). Among these deaths, 84% are due to heart attack and stroke.

Cardiovascular risks are multifactorial and heterogeneous, and therefore complex to manage and model from a (re)insurance perspective. As treatment of CVDs has much improved over the last 20 years, cardiovascular risks are evolving: today, people who have hypertension but carefully take their medications can be charged the standard life insurance premium if their blood pressure level gets back to normal. This would not have been possible a few years ago. Therefore, what used to be a risk yesterday might not be a risk today. As a consequence, today's risks are not necessarily tomorrow's risks. And the most innovative and inclusive insurance solutions need to reflect this.

SCOR wants to maximise the number of people being accepted by the insurance market with the optimal premium. With this objective in mind, SCOR Global Life launched Vitae Cardio in January 2021. Currently embedded in our Underwriting Manual Solem, the Vitae Cardio algorithm was built using machine learning techniques and takes into account new risk factors such as calcium score, waist circumference and physical activity, resulting in a more precise cardiovascular risk assessment and fairer pricing decisions.

Building a cardiovascular risk calculator such as Vitae Cardio requires deep knowledge of those risks and related operational challenges

When developing a framework and tool to assess cardiovascular risks from a medical underwriting perspective, the first step is to identify which medical and non-medical factors influence those risks. In the case of CVDs, medical research shows that the following factors play the most important role (which are not always independent from one another):

• **Age** – For people older than 65, CVD is the most important cause of death (40% of all deaths). Ageing is indeed often associated with a high prevalence of diabetes and high blood pressure, these are two key CVD risk factors (see below).

• **Gender** – Even if the difference has been decreasing over the last years, CVD and stroke risks currently remain higher for men than for women.

• **Tobacco** – Smoking alone accounts for approximately one in four deaths from CVD. The risk of developing CVD increases with the number of cigarettes smoked per day, and the duration of smoking.

• **Diabetes** – According to the American Heart Association, adults with diabetes are 2 to 4 times more likely to die from heart disease than adults without diabetes.

• **Blood pressure** – Hypertension is quantitatively the most important risk factor for premature CVD.

• **Dyslipidemia** – An abnormal amount of lipids in the blood can be a strong risk factor for developing CVD. However, the composition of cholesterol is more important than its total volume. While the LDL-C (Low-Density Lipoprotein Cholesterol) participates in vascular occlusion and is therefore a strong CVD risk factor, HDL-C (High-Density Lipoprotein Cholesterol) has a protective effect.

• **Personal CVDs history** – Despite having stable health and good treatment, a patient that has already suffered from CVD is at a high risk of developing new events.

• **Morphology** – The higher the Body Mass Index (BMI), the higher the risk of CVD (mainly because of an excessive body fat). Underweight people are also at higher risk of CVD. As a result, the relation between BMI and mortality is a "J-shaped" curve. Waist circumference can likewise be used to assess CVD risk from a morphologic point of view.

• **Calcium score** – Through a picture of the coronary artery measured by Computed Tomography (CT) scan of the heart, the calcium score allows to assess the risk of atherosclerosis or CVD. It is one of the most predictive single cardiovascular risk markers in asymptomatic individuals.

• **Physical activity** – Moderate physical activity is associated with lower cardiovascular risks, as it contributes to decreasing the heart workload and impacts metabolic changes (e.g., blood pressure, insulin sensitivity)

But using all the above factors to build an accurate and innovative cardiovascular risk calculator for medical underwriting purposes is not an easy task. Some constraints and challenges need to be considered:

• **Business and Underwriting Considerations** – Underwriters can have many cases to assess within a day. The number of variables and risk factors considered must be broad enough for robust risk assessment but not too complicated, as it can be challenging from an operations perspective. When medical underwriting is conducted manually (as is often the case for nonstandard risks), it is overly time consuming for each underwriter to enter 30 variables to obtain a quote. It also increases the operational risk, with a higher probability of input error.

• **Availability of the information** – Medical reports generally include standard information such as height, weight, blood pressure, etc. The factors used to assess cardiovascular risks need to be easily available to the underwriter to ensure they can be collected.

• **Database, modelling and medical research** – While the above risk factors are well studied and documented in medical research, a comprehensive and clear database is needed to lay out and model potential correlations between those factors and assess the risk accurately. Furthermore, the model needs to be flexible enough to gradually incorporate additional/ emerging factors when they are identified by medical research (e.g. air pollution).

Leveraging the NHANES database, Vitae Cardio's model has been built using machine learning techniques

WHAT IS THE NHANES DATABASE?

Vitae Cardio leverages the NHANES database, which is publicly available. NHANES (National Health and Nutrition Examination Survey) is a program of studies designed to assess the health and nutritional status of the US population, through representative surveys. It is supported by the Centre for Disease Control and Prevention (CDC) and is widely acknowledged as an official survey which collects accurate and reliable data. The survey is unique in that it combines interviews and medical examinations. This program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics.

The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. Furthermore, although this is a US database, some epidemiological studies have long shown that relative mortality risk remains consistent across geographies. Therefore, **we singled out NHANES as a valid data source to determine the risk factors for cardiovascular diseases and build our model.**

WHAT IS THE MODEL USED BY VITAE CARDIO FOR MORTALITY ESTIMATION?

Having a comprehensive database is just a first step. In order to assess the conditional mortality linked to an individual profile, a model needs to be built. First step was to transform continuous variables (the NHANES database keeps on record the medical and non-medical information of each individual included in the program) into categorical variables, more suitable for Underwriting purposes (although patient history matters, what matters even more is the current view of the medical information, such as current weight or blood pressure). Subsequently, four innovative survival analysis models were assessed to find out how reliable those models were to estimate extra all-cause mortality of the population with cardiovascular risk:

1. Poisson model

This model is based on proportional hazard rate assumption and piece-wise constant hazard assumption. The response variable is the yearly individual death number which is assumed to follow the Poisson distribution. Exposure time in each follow-up year is used to offset term to incorporate censoring information.

2. Logistic model with IPCW

This model is based on piece-wise constant hazard assumption and takes yearly death status as binary response variable. It changes the mortality estimation to a classification problem. To incorporate the censoring information, it uses inverse probability censoring weights (IPCW) principle which corrects the sampling weights to adjust the censoring problem.

3. Random survival forest (RSF)

This machine learning algorithm is dedicated to survival analysis. It takes follow-up time and death status as response binary group. It allows to solve the potential non-linear relationship between the response and the explanatory variables.

4. Gradient boosting machine (GBM)

Inspired by the Poisson model, this machine learning algorithm aims to improve the Poisson counting tree using boosting. The response variable and the offset term are the same as those in the Poisson model. Monotonicity constraints are imposed on explanatory variables to respect the constraint associated to medical literature.

In order to compare the performance and reliability of those models, two metrics have been used:

• **Concordance index (C-index)**, used to evaluate the predictions made by an algorithm, defined as the proportion of concordant pairs (i.e., proportion of accurate predictions) divided by the total number of possible evaluation pairs. Simply put, the concordance index assesses the proportion of cases where the model is right.

• **Integrated Brier Score (IBS)** is an overall measure of the predictions of the model at all times. In concrete terms, contrary to the concordance index, that assesses the ability of the model to discriminate efficiently, the Brier score gives an indication on "how close to reality the estimation predicted by the model is across time and what is the error margin".

While IBS is a valuable metric, C-index is the metric we consider to be the most important in our model. Beyond those metrics, two additional tests have been conducted:

• **Consistency test** – this test aims to estimate whether the model developed makes sense and shows results are aligned with medical research across 5 risk factors (smoking status, diabetes, blood pressure, cholesterol rate, BMI). For example, a model can only be efficient and selected if the loading for a non-smoker is always lower than the loading for a smoker.

• **Profitability test** – while from a statistical perspective the accuracy can be estimated by C-index and IBS, from an underwriting and actuarial perspective, the loss ratio (defined as the ratio of total technical losses divided by premiums) needs to be assessed, focusing on the substandard risks assessed through the tool developed rather than on the general portfolio.

For each of the tests and metrics above, we then ranked the four different models assessed from 1 (best performance) to 4 (worst performance), except for the consistency test (for which assessment is binary: Pass or Fail). Results are as follows:

Model	C-index	IBS	Consistency	Profitability
Poisson	3	3	P	2
Logistic	2	1	P	3
RSF	4	4	F	4
GBM★	1	2	P	1

★ **We therefore selected the GBM model, which stood out as the most accurate and robust model.**

Conclusion

By design, Vitae Cardio will make it possible to propose a finer rating to a larger insurable population and will therefore contribute to a more inclusive underwriting. Data allows to do this, thanks to the new risk factors it permits to explore, providing a lot of value in the underwriting process. But data alone was not enough to build Vitae Cardio. Vitae Cardio was built with Knowledge. The type of Knowledge that turns raw data into meaningful and relevant input, ultimately allowing to extend the life insurance safety net for previously underserved populations.