



First Prize Winner

Improve Insurance Accessibility via Artificial Intelligence

Kailan Shang, FSA, CFA, PRM, SCJP

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

When underwriting a life insurance application, a risk rating is usually assigned depending on demographic, medical, and economic information of the insured. It is possible that a higher-than-expected risk rating is determined based on collected information that leads to a rejection of the application or an unaffordable premium rate. The overestimation may be caused by limited data used in the analysis. Alternatively, traditional rate setting algorithms may not be able to easily handle additional factors that may justify a lower risk rating. With the aid of artificial intelligence (AI) models and improved data availability, more refined underwriting and pricing processes can be developed to improve accessibility and affordability of insurance. In essence, for individuals with a high risk rating, AI can be used to better assess if the risk rating is indeed that high, and further categorize them into more granular risk levels that will make an economic difference. In the following discussions, our focus is on mortality risk assessment, acknowledging much wider applications of AI in other risk types.

DATA

Life insurance carriers collect a variety of information to assess the mortality risk, with same examples listed in Table 1 under date type "Traditional underwriting/pricing factors."

Table 1
EXAMPLES OF DATA FOR MORTALITY RISK ASSESSMENT

Data Type	Data Category	Example
Traditional underwriting/ pricing factors	Demographic information	<ul style="list-style-type: none"> • Age • Gender • Residence • Race
	Medical information	<ul style="list-style-type: none"> • Height • Weight • Smoking • Use of drugs • Medical diagnosis • History of medical treatments
	Family medical history	<ul style="list-style-type: none"> • Anemia • Alzheimer’s disease • Asthma or other respiratory conditions • Brain disorders • Cancer • Diabetes • High blood pressure • High cholesterol levels • Immune deficiencies • Kidney, liver, or heart disease
	Occupation	<ul style="list-style-type: none"> • Occupation type • High risk occupation
	Lifestyle	<ul style="list-style-type: none"> • High risk activities
New underwriting/ pricing factors	Medical diagnostic data	<ul style="list-style-type: none"> • Age at diagnosis • Histological type • Insurance status • Involvement of lymph nodes • Primary tumor site • Reason for no surgery • Stage information • Surgery procedure • Tumor size and extension • Tumor type (Pos/Neg)
	Health data	Health monitoring data that may record heart rates, sleeping habits, physical activities. Instead of answers to a few questions, the data contains much more information to make individualized analysis of the implication on mortality risk.
	Image data	In addition to medical diagnostic data, medical image data may be used directly to perform AI-backed medical diagnosis and mortality risk assessment.
	Lifestyle data	Lifestyle and may be tracked and updated based on social media data.

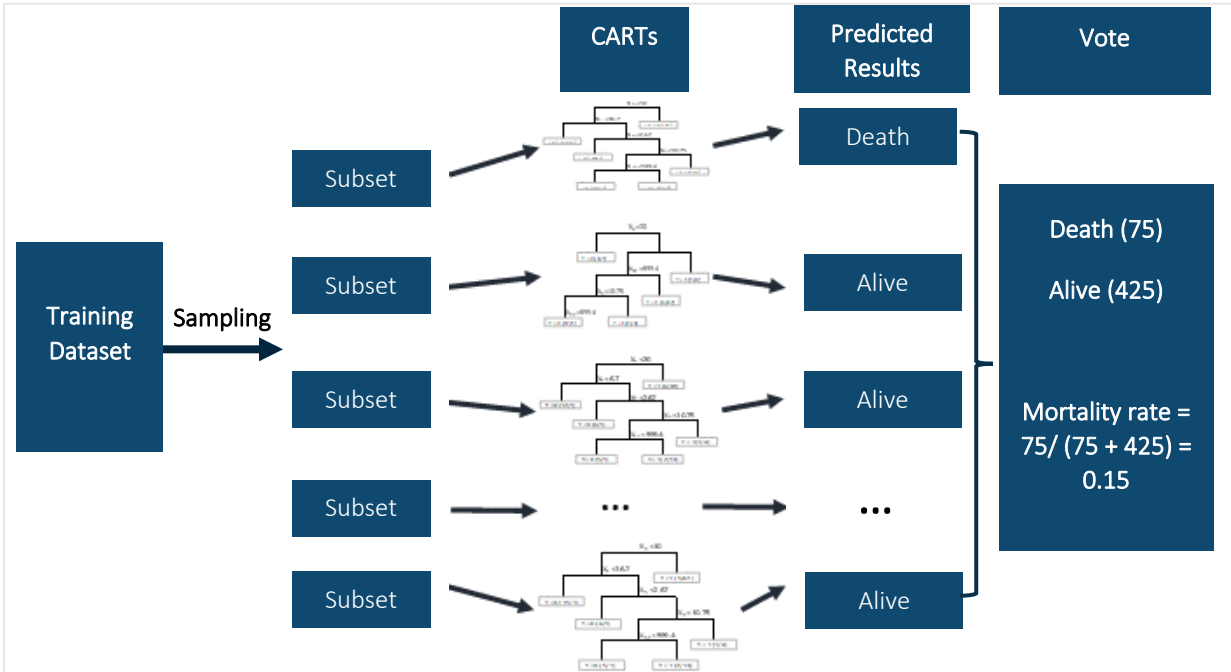
In addition, new factors such as detailed medical diagnostic data and health monitoring data may be used for refined mortality risk assessment with the support of AI models. Compared to answers to specific questions for the traditional factors, new factors usually contain more data in new formats that is difficult to use directly by underwriters or traditional pricing models. It requires AI models to analyze the new data and predict mortality risk of individuals in a more accurate way.

AI MODEL

In the field of AI, certain machine learning models can process large amounts of data and factors and use them in prediction. In the context of life insurance, AI models can be trained to establish the relationships between the mortality risk, or specifically the mortality rates, and the underwriting and pricing factors. Once the prediction accuracy is satisfactory, AI models can be used to automatically determine mortality risk at individual insurance coverage levels. Tree based models and deep learning models are promising candidates to improve mortality rate prediction.

Unlike traditional pricing models such as linear regression models or generalized linear models, **tree-based models** switch from formulas to decision rules for prediction. In a tree, leaves represent different subgroups and branches represent the rules to split into subgroups based on explanatory variables. The prediction is based on the value of the leaves that are in the same subgroup. Classification and Regression Tree (CART) models are a basic form of tree-based models. CART models build trees to split the data based on explanatory variables. At each split, a variable is used to separate the data into two subgroups. The variable is chosen to provide the best split that improves the purity of the data in the subgroups. More advanced tree-based models are built upon CART. The famous Random Forest models are a random version of the CART models. Multiple subsets are sampled from the training dataset which includes both traditional and new underwriting/pricing factors and observed mortality rates. Each subset is used to build a CART model. Explanatory variables are sampled as well so that the relationship between the mortality rate and the explanatory variables will not be dominated by the most important ones. Less important explanatory variables can contribute to the final prediction as well. Figure 1 illustrates the structure of a Random Forest model that may be used for mortality rate estimation. The final prediction is calculated as the average prediction by individual CART models.

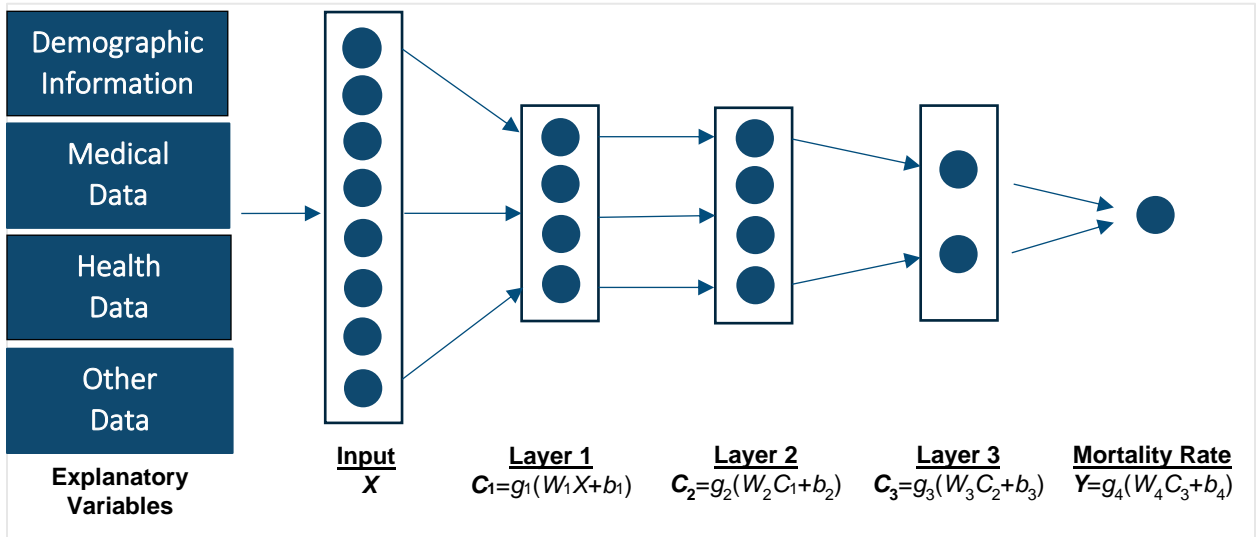
Figure 1
RANDOM FOREST MODEL STRUCTURE EXAMPLE



Gradient boosting machine (GBM) is another decision tree-based ensemble method. Each tree is a weak estimator trying to estimate the residual error that the estimation of previous trees has caused. Gradually with a sufficient number of decision trees, the estimation error will decline to a very low level. Unlike Random Forest models which use parallel trees to predict in aggregate, GBM is a sequential tree model.

Deep learning is a machine learning method that relies on artificial neural networks to represent the relationship between the response variable such as mortality rate and explanatory variables such as insurance pricing factors in an approximated way. “Deep” refers to multiple layers of neurons usually required in deep learning to be able to have a good approximation of the relationship. The most basic type of artificial neural network is fully connected neural networks, as illustrated in Figure 2.

Figure 2
ARTIFICIAL NEURAL NETWORK STRUCTURE EXAMPLE



C_i : the value of neurons in hidden layer i

g_i : the activation function for hidden layer i . g_4 is the function for the output layer. A common activation function is the sigmoid function $\frac{1}{1+e^{-x}}$ (a.k.a. logistic function). Many other activation functions are available as well.

The neural network can still be represented as a function of the input X . The function is a few linear layers ($WX+b$ and $WC+b$) and nonlinear layers (activation function) stacked together:

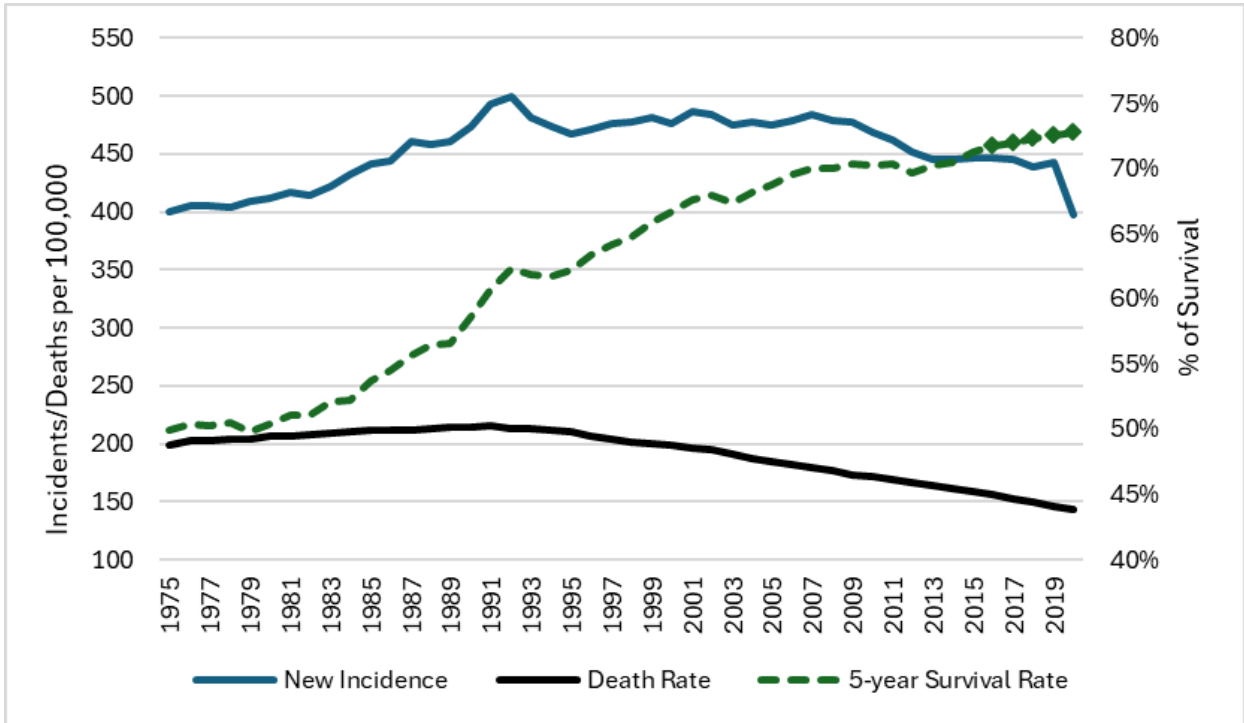
$$Y = f(X; W_1, b_1, W_2, b_2, W_3, b_3, W_4, b_4)$$

It is also proven that deep learning models such as recurrent neural networks and convolutional neural networks are good at utilizing text data and image data in prediction. Mortality risk assessment can potentially benefit from utilizing these nontraditional data types.

EXAMPLE: CANCER PATIENT MORTALITY PREDICTION

In general, it is difficult for cancer patients, whether they are under treatment or in recovery, to get individual life insurance coverages. High mortality risk leads to a low approval rate of life insurance applications. Even approved, the premium rate can be unaffordable. On the other hand, cancer incidence rate is quite stable while the overall mortality rate has been decreasing and more patients have a longer survival period. Better early detection, prevention, and medical treatment are likely to lead to lower mortality rates and longer survival periods, with the trend observed in Figure 3. Death rate has consistently dropped in the past three decades and the 5-year survival rate has increased as well.

Figure 3
U.S. CANCER INCIDENCE RATE, DEATH RATE AND SURVIVORSHIP (1975 - 2020)



Notes:

Data Source: National Cancer Institute, NIH, DHHS, "Cancer Trends Progress Report."

5-year Survival Rate: the rates of year 2016 to year 2020 are estimated by National Center for Health Statistics.

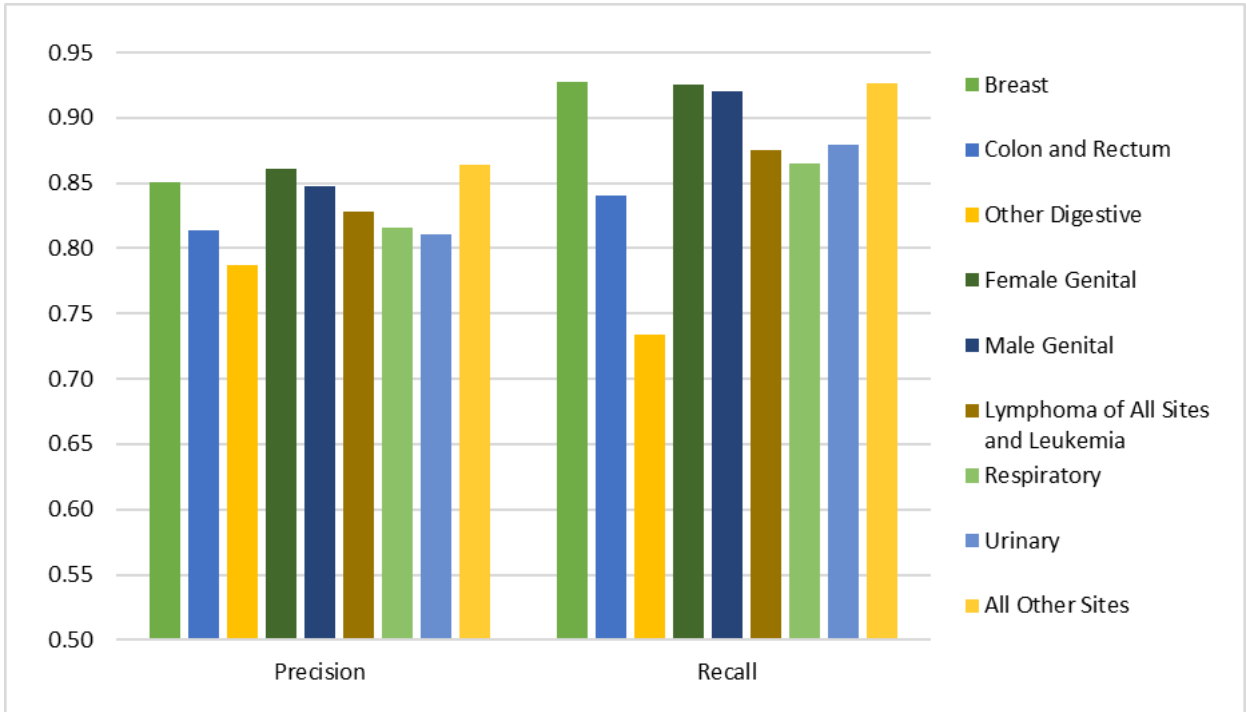
To improve cancer patients' and survivors' access to life insurance, it is necessary to evaluate the mortality risk of a cancer patient on an individual basis with both new data and new models. To illustrate its feasibility, medical treatment and diagnostic data¹ is used to predict not only the chance of survival but also the term structure of mortality. For life insurance products such as a term life product, the term structure of mortality affects not only the amount but also the timing of death benefit and premium income.

Using both tree-based models and deep learning models, a reasonably high level of accuracy of mortality rate estimation can be achieved. The precision and recall of the most accurate models among cancer types in this example is illustrated in Figure 4. Precision measures the Type I error² and recall measures the Type II error. The model accuracy based on validation results varies by cancer type.

¹ The Surveillance, Epidemiology, and End Results (SEER) research data is used in the example. The dataset includes U.S. cancer incidence and population data including age, gender, race, year of diagnosis, geographic location, tumor size, tumor location, tumor type (benign or malignant), histological information, diagnostic result, medical treatment, survival period, etc. Details can be found at <https://seer.cancer.gov/>.

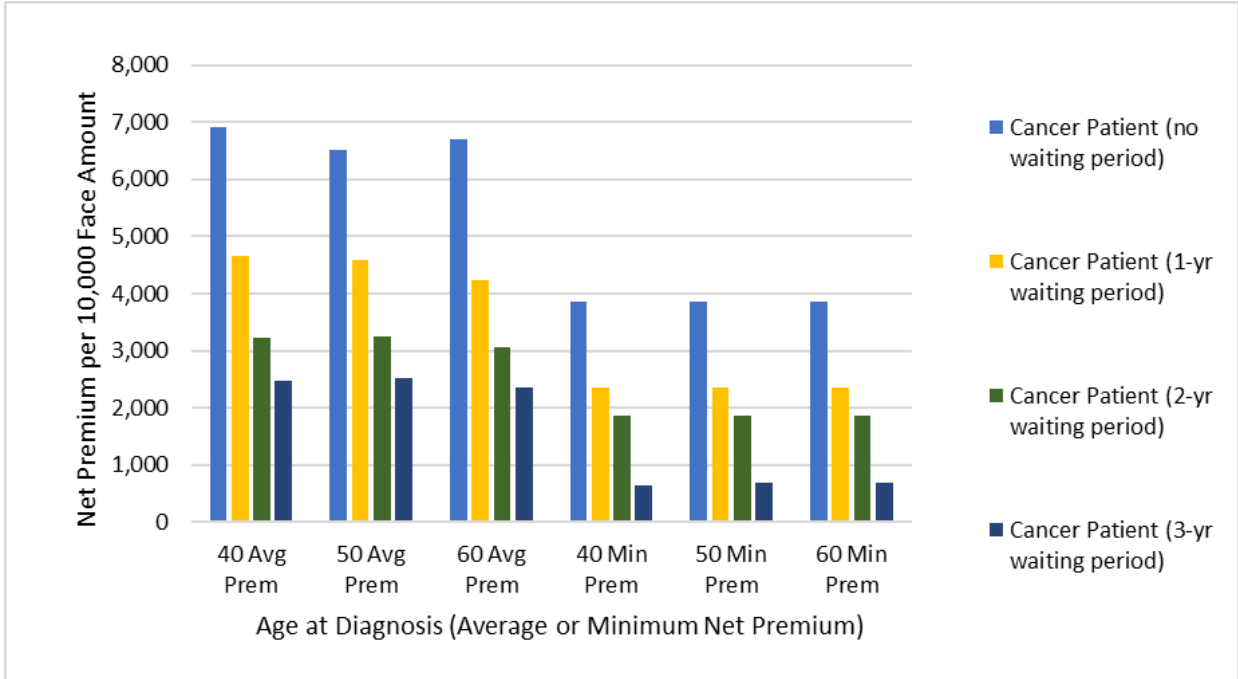
² Recall from classical statistics, a Type I error is a false positive where you reject a true hypothesis. A Type II error is a false negative and occurs when you fail to reject a false hypothesis.

Figure 4
AI MODEL PERFORMANCE EXAMPLE



Using the AI models, mortality rate prediction can be made for individual cancer patients. Premium rate of life insurance products can be calculated and compared among individuals. Figure 5 illustrates the net single premium of a 10-year term life product for some sample breast cancer patients, diagnosed at age 40, 50, and 60. Among the sample patients, both the average premium and the minimum premium (lowest predicted mortality rates) are shown.

Figure 5
SAMPLE 10-YEAR TERM INSURANCE NET SINGLE PREMIUMS BASED ON AI MODEL



Relatively low risk patients can be identified, and the impact is economically meaningful based on the difference between average and minimum premium amount. With more accurate underwriting and pricing for individual patients, it could lead to a higher acceptance rate of insurance applications from cancer patients and lower premium rates for applicants with low risk.

This example only uses part of the available data to predict the mortality rate of cancer patients. Other data such as computed tomography scan of tumors may be used to further improve the accuracy of prediction.

PREDICTION ERROR QUANTIFICATION

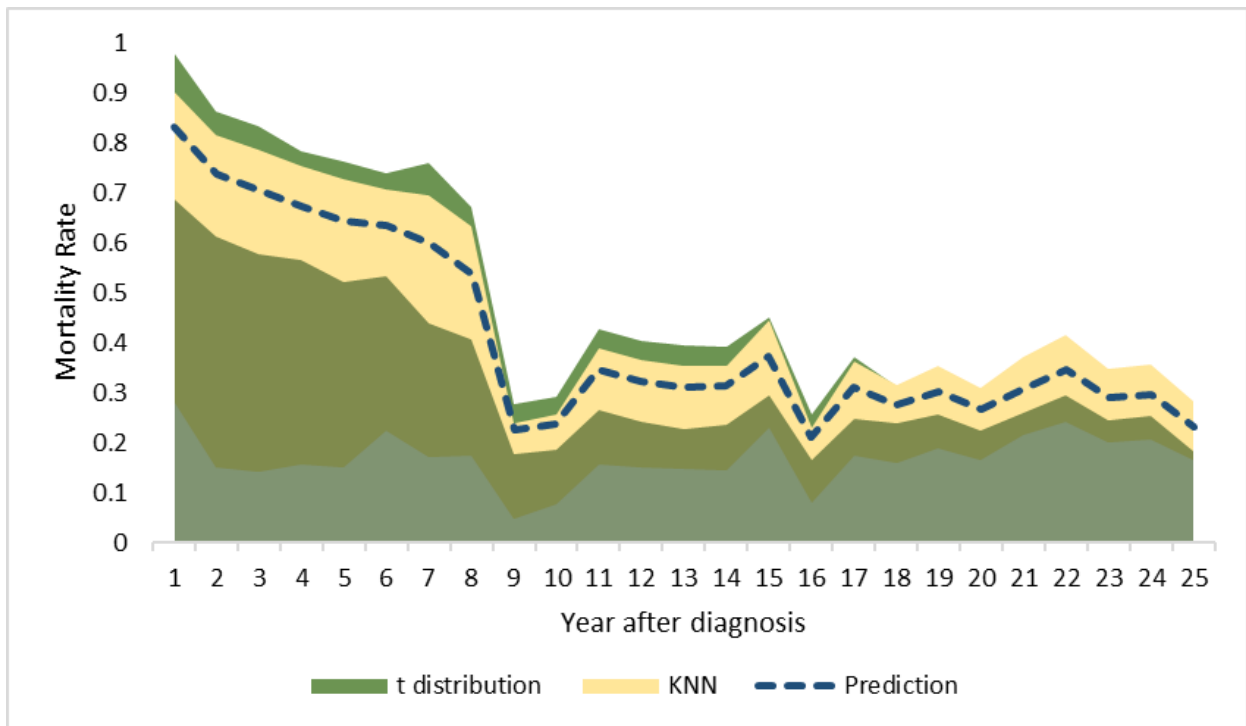
While AI models can predict mortality rates, the uncertainty of prediction needs to be quantified for mortality risk assessment and insurance pricing. However, for many predictive models, a prediction interval is not mathematically tractable. Nonparametric methods can be used to estimate the interval using the results of similar cancer patients. To estimate the prediction interval of a cancer patient’s mortality rates, similar cases in terms of explanatory variables can be searched. By finding the most similar cases (nearest neighbors), the variance of their estimates can be used to derive the prediction interval in a practical way. This can be realized by applying the k nearest neighbors (KNN) model, with the prediction interval constructed as follows.

- $$\left(\hat{y} - t_{k-1} \left(\frac{1-CL}{2} \right) s_k, \hat{y} + t_{k-1} \left(\frac{1+CL}{2} \right) s_k \right)$$

\hat{y} : predicted mortality rate.
 t_{k-1} : Student t value at the desired confidence level (CL) with a degree of freedom equals $k-1$.
 s_k : sample standard deviation = $\sqrt{\frac{\sum_{i=1}^k (y_i - \bar{y})^2}{k-1}}$ with y_i as the prediction for neighbor i and \bar{y} as the average prediction of the k neighbors.
- $$\left(\hat{y} - \left(\bar{y} - \left(\frac{1-CL}{2} \times 100 \right) th\ percentile \right), \hat{y} + \left(\left(\frac{1+CL}{2} \times 100 \right) th\ percentile - \bar{y} \right) \right)$$

The first approach assumes a Student t distribution and the second approach has no distribution assumption but uses empirical experience to determine the interval. The number of nearest neighbors to be sought can be set based on experience or determined by a maximum distance threshold. Figure 6 illustrates the mortality rate prediction interval of a breast cancer patient diagnosed at age 40 with a confidence level of 95%, using both approaches. For the approach using KNN, 100 nearest neighbors are used to construct the confidence interval. The interval is quite large during the first few years.

Figure 6
SAMPLE PREDICTION INTERVAL OF CANCER PATIENT MORTALITY RATES



Prediction intervals can be used to measure the robustness and credibility of individual mortality rate prediction. The upper bounds of the intervals may be more important than the estimation itself when making insurance underwriting and pricing decisions. Stress scenarios can be constructed accordingly to ensure that retained mortality risks do not exceed any risk limit. The ability to quantify prediction errors is instrumental to AI implementation in terms of getting stakeholder buy-ins and comprehensive risk assessment.

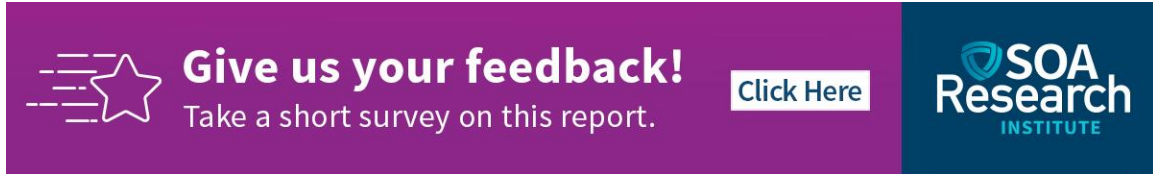
CONCLUSION


For high-risk life insurance applicants who may be rejected or given a high insurance quote, their risk may be overestimated by traditional data and models. AI models can help incorporate more individual factors into insurance underwriting and pricing processes in a data-driven and automated way. Non-traditional medical diagnostic data, health data, and lifestyle data are potentially new data types that can improve mortality risk assessment. Advanced AI models such as tree-based models and deep learning models may be adopted to analyze new data types that are usually more complex and of larger volume. In an example of predicting mortality rates of individual cancer patients, it is illustrated that the improvement due to AI models is financially meaningful and can potentially lead to a higher level of access to insurance. At the same time, prediction errors of AI models cannot be ignored and ideally can be quantified to ensure a robust implementation of AI in mortality risk assessment.

AI models may bring other issues that need to be addressed for a successful implementation. For example, new data may not be allowed to be used due to data privacy issues. AI models may unconsciously bring in biases to certain subgroups due to data limitation and not well-designed algorithm. Feasibility study needs to consider these issues as well, in addition to the potential gain from adopting AI models.

* * * * *

Kailan Shang, FSA, CFA, PRM, SCJP is a Director at Aon PathWise. He can be reached at klshang81@gmail.com.



 **Give us your feedback!**
Take a short survey on this report. [Click Here](#) 