Predictive Modeling Techniques Applied to Quantifying Mortality Risk

By Vincent J. Granieri

Actuaries are familiar with the interaction of art and science in their work. Some view underwriting in the same way, perhaps concluding that underwriting leans more toward art than science. With the advent of powerful computers and predictive modeling tools, it is possible to analyze survival data and produce statistically credible underwriting models that predict relative mortality risk among individuals based on demographic information and relevant conditions. In this article, we will discuss the use of the Cox proportional hazards model in developing a predictive underwriting model that produces a mortality multiplier for each individual. This multiplier can serve as the basis for debits and/or credits as it expresses the relative risk of having a given condition vis-à-vis not having it.

Further, we will attempt to quantify the impact on survival, if any, of being a member of certain subpopulations. We were looking to validate the time-accepted concepts of the wealth effect (in the wealthier subpopulations, which is beyond the scope of this paper) and antiselection (among insureds who sell their policies) in our population.

COX PROPORTIONAL HAZARDS MODEL

The Cox proportional hazards model was introduced in 1972 as a method to examine the relationship between survival (mortality) and one or more independent variables, called explanatory variables. Some advantages of the Cox model are that it can utilize many underwritings on the same life and can handle data that is right censored, i.e., subjects can leave the study at any time or the study can end before all subjects have died. The Cox model does not require knowledge of the underlying (base) survival curve, which is advantageous; however, we will see that this advantage also brings challenges when analyzing mortality.

Cox model results are expressed as the logarithm of the hazard so technically, the relative risk factor for each variable is obtained by raising e to the power of the log(hazard). Actuaries will recognize this as consistent with Gompertz. The relative risk factor is interpreted just as it sounds: It describes the force of mortality of subjects having a certain condition relative to that of the reference population, who do not have that condition. A relative risk factor of two for a condition means the subject is twice as likely to die as another subject who does not have that condition.

As an aside, we utilized the survival package in the R statistical language to produce our survival models. It is particularly well-suited for this type of analysis. Other popular statistics programs, such as SAS, also contain survival models using the Cox model.

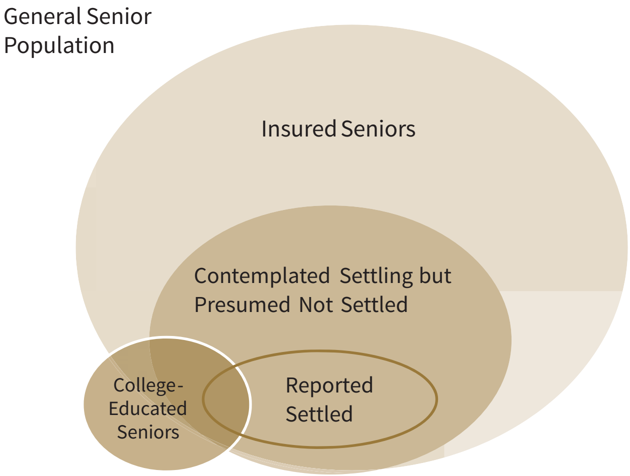
THE ISSUES

A most important issue was that of the underlying mortality distribution. We already had produced mortality tables that varied by age/gender/tobacco use. What then should we do with the Cox model results that also calculated the impact of these variables? It was also very important to ensure that the explanatory variables were truly independent. If not, spurious results would ensue. We also had to redefine certain variables, such as body mass index (BMI), where the risk was actually related to straying from the ideal BMI measurement, rather than the measurement itself. There were many other issues, too numerous to mention in an article of this length.

INPUT DATA

For this exercise, we had available to us over 200,000 underwriting events on 80,000+ unique senior lives, which took place over a 15-year period, primarily in the life settlement market. Figure 1 is a graphic description of the major subpopulations of the universe of senior lives and the populations we studied. At the highest level, there is the general senior population. Some of these seniors have purchased insurance, creating a subpopulation, which can be further broken into two subpopulations:

Figure 1: Senior Populations


General Senior Population
Insured Seniors
Contemplated Settling but Presumed Not Settled
College-Educated Seniors
Reported Settled

> The most important conclusion we drew from this exercise was that despite our best efforts to quantify every aspect of underwriting, there is still considerable judgment brought to bear in the process.

those who actually sold their policies on the secondary market and those who contemplated such a sale but, for some reason, did not conclude the sale. These latter two subpopulations were the basis for our study of antiselection. There is also a small population of college-educated seniors, some of whom can also be associated with the other populations above, which formed the basis for our study of the wealth effect. This data included demographic information such as age, gender, date of birth and date of death. It also included various underwriting conditions such as BMI, smoking status and indicators for various diseases. Included were favorable conditions, such as family history of longevity (parents/siblings who lived beyond age 85) and good exercise tolerance.

CREATING COX PROPORTIONAL HAZARDS MODELS

There was significant data preparation involved. We set up the reference population, which we chose to be males who were age-appropriately active, who did not sell their policies and did not use tobacco. Variables were determined to be either continuous (age, BMI), where the condition has infinite possible values, or binary (coronary artery disease, osteoporosis), where the condition either exists or does not. This required considerable judgment and depended on the availability and form of the data.

Once the data were prepared, we began the process of determining which conditions were statistically significant in predicting mortality. We underwent an iterative process. The Cox models were run with every variable included at first. Then we reran the models, first eliminating most of those variables with a p-value greater than 0.2. This means we were excluding those conditions where the probability that the relative risk shown was due to random fluctuation was over 20 percent. These models were again rerun, this time eliminating those conditions with a p-value greater than 0.1. Finally, we reran the models, including only those conditions where the p-value was at most 0.05.

RESULTS

Figure 2 represents only a portion of the output from our models, consisting of conditions that were included in all runs even if they did not meet the criteria for continued inclusion above. As we advanced through the process, we felt strongly these were fundamental variables that clearly impacted survival and should be included in the analysis regardless of their p-values. In reality, only one variable (rare smoker) would have been eliminated, presumably due to data scarcity. There were a number of other explanatory variables that also made the final cut, but space does not allow their inclusion herein.

Pink/green shading indicates that a condition is hazardous/protective, with the 95 percent confidence limits and p-values also shown. For example, the female hazard is 0.694 of that of males (1.0, as males are the reference). Therefore, the female mortality rate is found by multiplying the male rate by 0.694 for all ages. The hazard for age is 1.08, which means that for any age, the mortality rate for the next higher age is found by multiplying the mortality rate of the first age by 1.08. The smoker hazard is 1.887 times that of the reference, which is nonsmokers; it follows that the smoker mortality rate then is 1.887 times the corresponding nonsmoker rate. This is where the disadvantages of the Cox model came into play. The issue became whether we should replace our base tables for male/female, smoker/nonsmoker with tables based only on the proportional hazards produced in our predictive models and our base male nonsmoker table. After reviewing the model results for consistency with them, we decided to use all four of our existing base tables; however, we broke out antiselection explicitly.

Figure 2

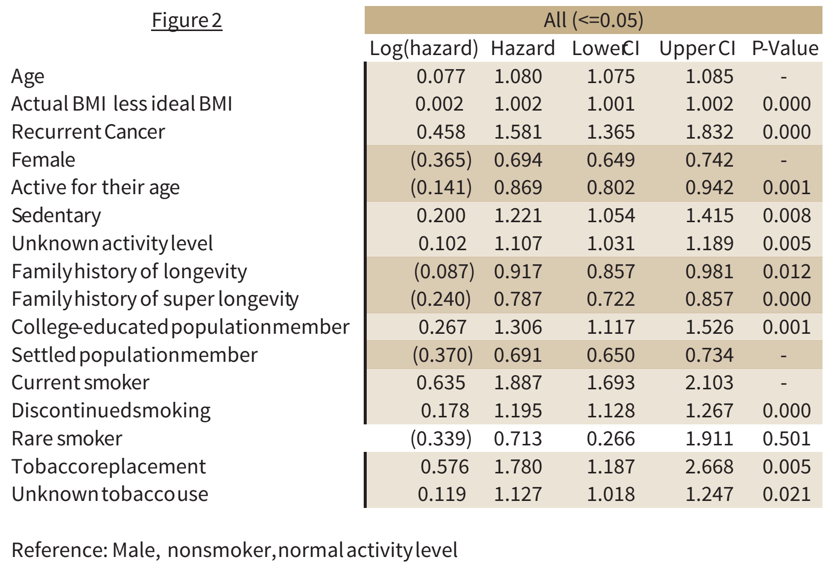


Figure 2	All (<=0.05)				
	Log(hazard)	Hazard	Lower CI	Upper CI	P-Value
Age	0.077	1.080	1.075	1.085	-
Actual BMI less ideal BMI	0.002	1.002	1.001	1.002	0.000
Recurrent Cancer	0.458	1.581	1.365	1.832	0.000
Female	(0.365)	0.694	0.649	0.742	-
Active for their age	(0.141)	0.869	0.802	0.942	0.001
Sedentary	0.200	1.221	1.054	1.415	0.008
Unknown activity level	0.102	1.107	1.031	1.189	0.005
Family history of longevity	(0.087)	0.917	0.857	0.981	0.012
Family history of super longevity	(0.240)	0.787	0.722	0.857	0.000
College-educated population member	0.267	1.306	1.117	1.526	0.001
Settled population member	(0.370)	0.691	0.650	0.734	-
Current smoker	0.635	1.887	1.693	2.103	-
Discontinued smoking	0.178	1.195	1.128	1.267	0.000
Rare smoker	(0.339)	0.713	0.266	1.911	0.501
Tobacco replacement	0.576	1.780	1.187	2.668	0.005
Unknown tobacco use	0.119	1.127	1.018	1.247	0.021

Reference: Male, nonsmoker, normal activity level

CONCLUSIONS

The most important conclusion we drew from this exercise was that despite our best efforts to quantify every aspect of underwriting, there is still considerable judgment brought to bear in the process. However, there is also much useful information that predictive models can provide us because of their ability to process large amounts of data quickly and efficiently. We did validate the antiselection that occurs between those who actually sell their policy versus those who do not (as seen by the hazard ratio of 0.691 for the settled population members in Figure 2). Some results confirmed our clinical judgment; for example, an active lifestyle or family history of longevity are indicators of higher survival rates. Other things went against our clinical judgment; for example, cardiac-related conditions, while still hazardous, were no longer as significant as we thought.

Then there were the confounding results. Hyperlipidemia (high cholesterol) was shown to be protective. We attributed this to the ubiquity of statins. There were a number of other conditions shown to be mildly protective, things such as benign prostatic hyperplasia (BPH), sleep apnea, use of blood thinners and benign colon polyps. We concluded that these were indicators of frequent/better quality of health care, which would allow for early detection and mitigation of more serious risks. Similarly, family history of heart disease and cancer were seen as mildly protective, presumably due to their providing early warning signals to take protective actions, such as better diet and more exercise in the case of heart disease and more frequent screenings in the case of cancers.

BUSINESS OUTCOMES

This analysis was the basis for changes in our debit/credit underwriting model. We replaced an additive model based only on clinical judgment with one that was exponential in nature, which provided more consistency to mortality research. The new model was quite flexible and allowed us to continue to factor in clinical judgment where appropriate. For example, we used the relative risk factor for smokers who quit, but isolated the impact by time since smoking ceased, reducing the debit as time went on. ■

Vincent J. Granieri, FSA, EA, MAAA, is chief executive officer at Predictive Resources LLC. He can be reached at *vgranieri@predictiveresources.com.*