LifeScoreLabs℠

Improving the accuracy and transparency of underwriting with artificial intelligence to transform the life-insurance industry

Leveraging a data set spanning 20 years of applications at MassMutual, this technical paper describes the development of an industry-leading model that estimates mortality risk, known today as LifeScore Med360.

September, 2020

Improving the Accuracy and Transparency of Underwriting with Artificial Intelligence to Transform the Life-Insurance Industry

Marc Maier, Hayley Carlotto, Sara Saperstein,
Freddie Sanchez, Sherriff Balogun, Sears Merritt

■ *Life insurance provides trillions of dollars of financial security for hundreds of millions of individuals and families worldwide. To simultaneously offer affordable products while managing this financial ecosystem, life-insurance companies use an underwriting process to assess the mortality risk posed by individual applicants. Traditional underwriting is largely based on examining an applicant's health and behavioral profile. This manual process is incompatible with expectations of a rapid customer experience through digital capabilities. Fortunately, the availability of large historical data sets and the emergence of new data sources provide an unprecedented opportunity for artificial intelligence to transform underwriting in the life-insurance industry with standard measures of mortality risk. We combined one of the largest application data sets in the industry with a responsible artificial intelligence framework to develop a mortality model and life score. We describe how the life score serves as the primary risk-driving engine of deployed algorithmic underwriting systems and demonstrate its high level of accuracy, yielding a nine-percent reduction in claims within the healthiest pool of applicants. Additionally, we argue that, by embracing transparency, the industry can build consumer trust and respond to a dynamic regulatory environment focused on algorithmic decision-making. We present a consumer-facing tool that uses a state-of-the-art method for interpretable machine learning to offer transparency into the life score.*

Life insurance is a critical financial tool for millions of households, providing security to families by reducing the financial impact of an untimely death. In the United States alone, life-insurance companies collectively manage trillions of dollars of protection while annually disbursing billions of dollars to beneficiaries; according to the American Council of Life Insurers, at the end of 2018, there was nearly $12.1 trillion of active coverage for individuals and $57 billion in payments to their beneficiaries.[1] To support this large-scale financial ecosystem while simultaneously offering affordable prices, insurers must estimate the mortality risk of individual life-insurance applicants through an underwriting process. The accuracy of this underwriting ultimately drives the long-term stability of the life-insurance industry because the collective sum of incoming premiums, which are fixed post-underwriting, must be sufficient to offset future payouts from guaranteed death benefits.

Unlike most types of insurance that are renewed and reassessed annually (such as property and health), nearly all life-insurance policies are one-time, long-duration contractual agreements. Thus, the veracity and completeness of health and behavioral data used for mortality-risk assessment is paramount. For the past few decades, life underwriting has been guided by manual review and point-based systems that predominately consider factors independently. Consequently, traditional underwriting limits the degree to which insurers can accurately estimate risk from data and achieve optimal price efficiency of products.

  ISSN 0738-4602

The industry also faces systemic challenges beyond the technical complexity of estimating mortality risk. Existing processes for providing coverage have been slow to adapt to changing demographics and to meet expectations of a digitally enabled and rapid service customer experience found in nearly all commercial applications. According to market research, there is a vast population of uninsured and underinsured households, estimated at $25 trillion, that the industry has yet to reach (Abrokwah et al., 2018). And the availability of new data sources and the opaque usage of algorithms pervading our daily lives have led to increased privacy and transparency concerns by consumers, resulting in new regulatory frameworks, such as the California Consumer Privacy Act and the General Data Protection Regulation in Europe. We believe that an ethical and responsible use of artificial intelligence (AI) can solve some of these great challenges, broadly improving the financial wellness of society.

The availability of large historical data sets provides an unprecedented opportunity for AI to transform underwriting in the life-insurance industry. At MassMutual — a large insurance and financial services company — we have curated a data set of more than one million applicants spanning 20 years, representing one of the largest and most comprehensive application data sets in the industry. Leveraging this data, we develop an accurate, high-resolution mortality model[2] that generates a life score and serves as the primary risk-driving engine of algorithmic underwriting systems (Maier et al., 2019). Collaborating with actuaries, we design a novel evaluation framework to compare historical underwriting decisions against simulated model decisions and demonstrate that the life score outperforms traditional underwriting, yielding a nine-percent reduction in mortality within the healthiest pool of applicants.

While improving accuracy in underwriting through AI is a highly valuable endeavor, it is imperative to maintain privacy and embrace transparency in data and methodology. Building consumer trust through transparency and education will likely be a critical component for the industry to strengthen ties with existing markets and to reach historically underserved ones, especially with the rise of digital capabilities. We advocate for embedding state-of-the-art methods in interpretable machine learning, such as the Shapley Additive Explanation (SHAP) framework (Lundberg and Lee, 2017), that enable transparency in algorithmic scoring and decisions. To that end, we produce model explanations alongside each predicted life score to support human oversight during underwriting and offer explanations to applicants. We also introduce a simple, consumer-facing transparency tool, called MyLifeScore,[3] to demonstrate how various factors, such as blood pressure or family history, may drive individual risk, thereby helping to demystify the underwriting process.

Finally, we posit that a standardized life score underpinned by accuracy and transparency has potential for broader industry impact. A well-established life score that is understood by consumers has the ability to increase access to life-insurance protection. Additional benefits include enabling life-insurance-backed securities for a wholly new dimension in diversifying financial portfolios; expanding options for purchasing life insurance; and facilitating connections to personalized health and wellness programs.

Background

Life insurance is a financial tool that helps individuals secure the financial future of loved ones in the event of their passing. A life-insurance policy is an agreement whereby an insurer pays beneficiaries a sum at the time of a policyholder's death. In return, the policyholder pays premiums over a predefined period of time. Beneficiaries generally use the proceeds to pay for expenses that would have otherwise been paid for by the earnings of the insured.

Life-Insurance Underwriting

Most types of life insurance require an estimate of the expected lifetime of an individual at the time of application. This is referred to as *mortality risk*, and the process of collecting and analyzing data that describes such risk is known as *underwriting*. Actuaries compute the cost of covering mortality risk over the lifetime of the policy and translate it into a set of premium payments. The financial risk and general approval of the underwriting process is agreed upon with reinsurance companies — institutions that assume and further diversify a portion of the risk.

For the past few decades, life underwriting followed expert judgment aided by medical impairment manuals that specify guidelines for grouping individuals into broad classes of mortality risk. These manuals implement point-based systems, wherein debits and credits are drawn from medical studies that offer mortality ratios or survival probabilities for a given disease or impairment (Brackenridge, Croxson, and Mackenzie, 2006). Hundreds of medical and behavioral attributes, such as the presence of heart disease, use of statins to lower cholesterol, or family history of certain cancers, are mapped to point values and combined. A life underwriter reviews an application to complete the overall assessment, which then determines one of several discrete risk classes that drive premiums and are priced according to expected aggregate mortality.

Most large carriers are automating the logic of their underwriting manuals with rule systems. Automated rule systems provide fast and consistent decisions, but they perpetuate the simplified guidelines that they describe, leading to suboptimal risk-class assignments. For example, a laboratory test result that exceeds a prespecified threshold may disqualify an individual from certain risk class without considering related factors. Reflexive rules (that is, conditioning the threshold on age and sex) may classify risk more accurately; however, expanding the size and complexity of a rule system increases the technical burden that carriers must manage.

More recently, rich data sets and advancements in machine learning have enabled predictive models to improve mortality-risk classification in life underwriting. Predictive models that leverage survival analysis are grounded in the same theory behind traditional underwriting, but they can outperform human and rule-based methods by detecting complex, multivariate dependencies between health factors and mortality. This class of models is limited to estimating risk based on data attributes that have been collected for a large set of individuals with sufficient history to observe enough mortality outcomes. For data sources that have only recently been collected digitally, rule-based approaches and expert judgment may still be necessary to use in conjunction with models.

Predictive Models in Life Underwriting

The life-insurance industry is actively improving its digital capabilities and customer experience to narrow the uninsured and under-insured protection gap. This goal has led to the emergence of new data sources and underwriting paradigms that simplify and expedite the application process. In doing so, the industry needs mortality-risk scores to capture the predictive power of both traditional and nontraditional data sources to accelerate and improve underwriting decisions. While mortality modeling is a maturing topic of academic research and multiple vendors are developing novel solutions, adoption of predictive models in the industry is still in its infancy.

Acknowledging the value that AI can bring to life-insurance underwriting, researchers have built models to replicate historical underwriting decisions (for example, Boodhun and Jayabalan [2018]), improve aspects of the underwriting process (for example, policy routing; Dubey et al. [2018]), or in some cases, directly predict mortality risk. One carrier successfully applied fuzzy logic to codify underwriting guidelines and enable automation (Aggour et al., 2006). Generally, the approaches rely on regression or classification and use underwriting decisions as the target variable, as it is difficult to procure data with sufficient historical coverage to support survival models based on observed mortality. Such data sets are often proprietary or secured by institutions focused on data aggregation.

Data providers recognize the value of their large repositories and have developed predictive models based on different aspects of risk. Vendors that have long partnered with insurers to perform laboratory tests or supply prescription drug histories offer risk scores based on their clinical data. Other businesses that collect data sources not traditionally used in life underwriting, such as credit histories and public records, have built similar solutions that support accelerated underwriting. As electronic health records become more prevalent, vendors are attempting to embed this data into predictive models, and insurers are eager to incorporate reliable and accessible medical information.

Despite substantial research and an increasing number of vendor solutions, the industry has yet to capitalize on predictive models at scale. The use of machine learning to improve mortality-risk assessment is widely accepted as a necessary direction by industry leaders, but implementation has been difficult for an industry generally constrained by legacy technology. For the few cases achieving a successful implementation, the results have brought significant value to both carriers and customers. This highlights the opportunity and benefits that a standardized life score paired with an accompanying rules engine could provide to the broader industry.

Defining a Standard Life Score

If broadly adopted by life-insurance carriers, a standard, trusted life score that quantifies individual mortality risk would likely produce several benefits for the industry and beyond. Importantly, it could increase access to life-insurance protection by driving a marketplace for consumers to explore purchasing options given their fixed risk profile. The life score could also motivate carriers to connect their customers with wellness programs that aim to incentivize healthy behavior and quantify the benefit with prospective scores. Additionally, life-insurance–backed securities scored consistently and transparently could be traded in a capital market, creating a new mechanism to diversify financial portfolios. Acceptance of such a life score would hinge on industry and consumer trust. To this end, the algorithm generating the score must use well-understood and justified data inputs; the score should be transparent with respect to its contributing factors; and it should achieve state-of-the-art accuracy in predicting long-term mortality risk.

To justify use of data in a standard life score, there should be substantial causal and medical evidence tying attributes to mortality risk. Laboratory tests and health questionnaires have a longstanding precedent and medical basis for providing relevant information for assessing mortality risk in life-insurance underwriting. The score should also be simple to interpret by consumers, underwriters, and regulators. We accomplish this with a standard 0 to 100 scale, ranging from highest to lowest risk, corresponding to the health percentile of individuals relative to their peers. The score is exemplified as: If Carlos is a 55-year-old nonsmoking male with a life score of 87, he can be compared directly against and has lower mortality risk than Barry, another 55-year-old nonsmoking male with a score of 53. If Amy is a 35-year-old nonsmoking female with a score of 87, she does not necessarily present the same mortality risk as Carlos, but can be compared to other 35-year-old nonsmoking females.

Despite using well-understood inputs, a singular score does not offer transparency into its derivation. We compute the contribution of each health factor using a state-of-the-art method in model interpretability, such that the sum of the contributions equal the life score. Finally, to achieve reliability, the score must be based on comprehensive data, developed with a sound modeling methodology, and it must demonstrate highly accurate risk stratification. The following sections expand on these properties and the use of the life score at MassMutual.

Data

MassMutual has consolidated a digital record of applications for which a laboratory test was ordered during 1999 to 2019. Removing applications with a high degree of missing values, typically from incomplete applications, results in 1.5 million records with 13 million exposure years and 23,000 observed deaths. Of these applicants, thirty-nine percent are women and median ages are 41 and 39 years for men and women, respectively. The data set covers attributes drawn from laboratory tests and a lengthy health history questionnaire that accompanies the application process.

Laboratory Tests

Life underwriting typically includes a set of laboratory tests on blood and urine specimens. A vast medical and actuarial literature ties various tests directly with all-cause or specific causes of mortality, such as albumin levels (Goldwasser and Feldman, 1997). The laboratory test data provide exposure to a range of values across biophysical measurements such as build and blood pressure, lipids, liver function tests (for example, gamma-glutamyltransferase), kidney function tests (for example, creatinine), blood and urine proteins (including albumin, globulin, and microalbumin), blood sugars, and several indicators, such as cocaine and HIV.

Health History Questionnaires

Lab tests are an instantaneous view into individual health that yield substantial protective value for risk selection. Life underwriting also solicits information related to personal and family health history, as well as behavioral risk through an extensive questionnaire. Questions could include, for instance, whether an applicant has received a recent medical diagnosis of heart disease or endocrine disorder.

Partnering with a vendor specializing in handwriting recognition, we digitized the vast majority of MassMutual's paper and imaged archive. This endeavor was challenging due to a manual element of standardizing questions phrased differently across time, states, and product offerings. It also necessitated development of a complex data-ingestion pipeline to process the digitized responses. Despite the acquisition costs, these data enable a consistent mapping with current application questions, which are combined into a single data warehouse for modeling and analysis.

Health Trends across Time

Given the 20-year period of our data, we can observe trends in the distribution of certain laboratory values. For example, recent applicants exhibit lower cholesterol levels compared with those in earlier years, as shown in figure 1. This is consistent with medical research reporting similar trends over the same period (Rosinger et al., 2017). Variables that trend over time, referred to as covariate shift or nonstationarity, present a modeling challenge due to the temporal association with predictive variable. We apply a statistical adjustment that translates and controls for these temporal differences in distributions. With recent research discovering worsening mortality trends on specific subpopulations (Case and Deaton, 2015; albeit stemming from uncertain factors), it will be imperative to capture the changing dependence of laboratory tests and mortality risk.

Developing the Mortality Model

With medically justified inputs and state-of-the-art science, we leverage one of the largest and most comprehensive application data sets in the industry to develop a mortality model, validated with holistic criteria, to produce strong model performance in production.

Feature Selection

Feature selection was heavily influenced by medical and actuarial experts and validated with standard machine-learning techniques. The deployed mortality model relies on nearly sixty inputs captured in biophysical measurements, blood and urine specimens, and applicant health history questionnaires.

One challenge of application data is that questions and requirements are revised over time, often varying across states and product types. This requires fastidious mapping to produce consistent data for modeling. Given the recommendations of the medical team, we reviewed historical coverage of each variable to ensure alignment with documented underwriting and medical guidelines. We also assessed the statistical dependence with mortality inherent to each variable. For example, figure 2 shows how relative mortality varies by five-point bands of body mass index (BMI), exhibiting slightly elevated mortality for low BMI and steadily increasing mortality for higher ranges.

Improvements to feature generation and selection in subsequent model iterations involve regular, extensive reviews with MassMutual's medical team based on observations in model outcomes. This close partnership with experts is important for constructing an intuitive and medically relevant mortality model. Likewise, model results and updated medical knowledge assist in guiding changes to laboratory testing. For example, MassMutual recently shifted from using HbA1c — a measure of average blood sugar over a two- to three-month period and an important biomarker for diagnosing and monitoring diabetes — as a reflexive test to employing it as a screening test despite the increased cost. MassMutual historically tested HbA1c reflexively if certain conditions were met, such as a history of endocrine disorder or high glucose or fructosamine levels, which are cheaper, but less reliable markers of diabetes. Because HbA1c was partially missing in historic records due to it being a reflexive test, serum glucose and fructosamine were originally included in the model as the dominant

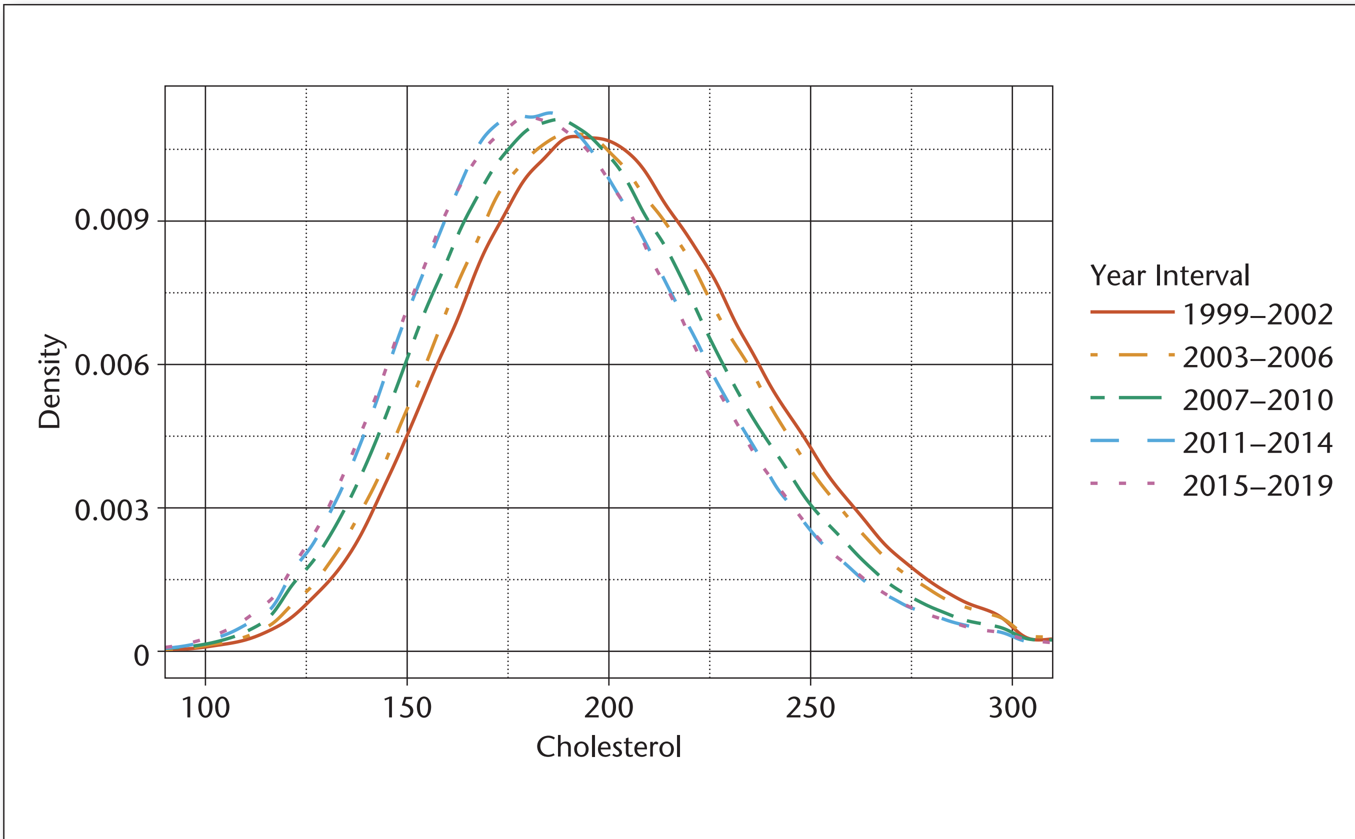

Figure 1. Grouping by Four-Year Bands, the Distribution of Cholesterol Trends Lower Over Time.

features that measure blood sugar. However, glucose is sensitive to fasting, and HbA1c is a strong measure of mortality risk (Stout et al., 2007). Therefore, we imputed HbA1c for historic cases by predicting its value from other variables and currently use HbA1c as the sole input for blood sugar, resulting in a more accurate model.

Aside from imputing historically reflexive tests, we have also imputed historical values of entirely missing variables that are consistently collected during the application process, such as pregnancy status. Successfully imputing historically missing values enables the model to support new and emergent inputs.

Model Development

The majority of predictive modeling tasks are classification-based (that is, estimating the probability of a discrete outcome) or regression-based (that is, estimating the expected value of a continuous outcome). In survival analysis, however, the outcome of interest is the duration until a binary event may occur. The primary goal of predictive modeling in the survival context — termed *survival modeling* — is to develop estimates of the survival, hazard, or cumulative hazard functions with respect to a set of observed covariates. The survival function, $S(t) = Pr(T > t)$, describes the probability that an event, occurring at random variable time T, occurs later than some given time t. The hazard rate,

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \le T < t + dt)}{dt \cdot S(t)},$$

is the rate of the event at time t conditioned on having survived until time t. The cumulative hazard function, defined as

$$\Lambda(t) = \int_0^t \lambda(u)\,du,$$

is related to the survival function as $\Lambda(t) = -\log S(t)$.

The Cox proportional hazards model is the most widely used statistical technique for estimating individual risk in studies of survival (Cox, 1972). It is a semiparametric regression model that assumes a linear functional form and proportional hazards for any two strata over time. In machine learning, the random forest method (Breiman, 2001) has been adapted by Ishwaran et al. (2008) to handle right-censored survival outcomes (called “random survival forests”), and efficient implementations exist (Wright and Ziegler, 2017). As a nonparametric, adaptive model, a random survival forest captures interactions and

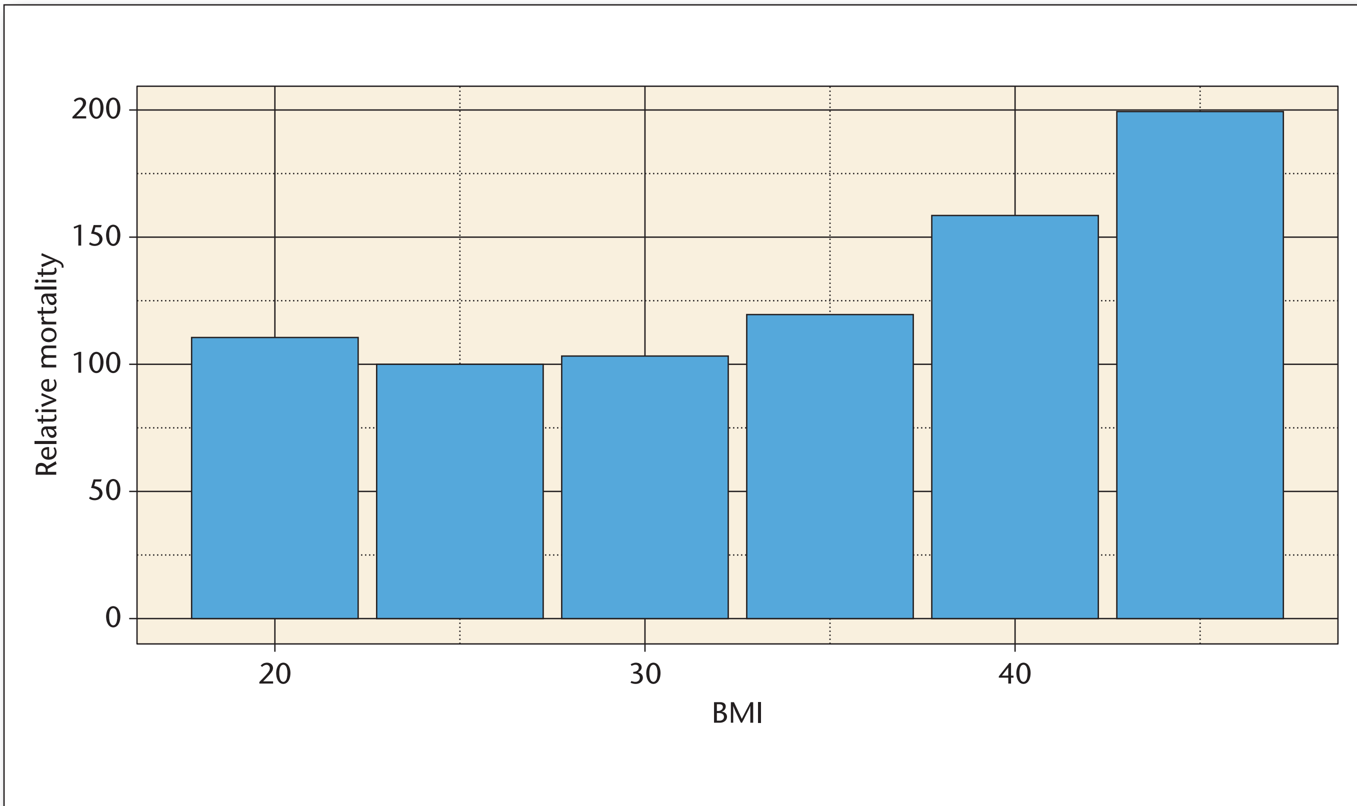

Figure 2. Trends in Aggregate Mortality Risk, Measured by Relative Mortality as a Function of Five-Point Bands of BMI.

nonlinear dependencies that are more subtle and complex than can be reflected by a linear model.

Emerging research aims to apply advanced statistical models to discrete-time survival analysis (Bender, Groll, and Scheipl, 2018) as well as survival extensions of deep learning (Katzman et al., 2016; Ranganath et al., 2016). However, scalable implementations are limited, with the most comprehensively developed survival suite existing in the R software environment. Thus, we focused our modeling on the random survival forest, as internal testing showed that it outperformed Cox proportional hazards models and a survival implementation of a deep neural network. Experiments iterated on findings drawn from our collaborative feature selection process, along with improvements through variable transformation, hyperparameter tuning, and sampling techniques. All experiments performed 10-fold cross validation, and held-out predictions were used to produce a suite of statistical, actuarial, and business-relevant evaluation metrics (detailed next).

The random survival forest mortality model directly estimates the cumulative hazard function, $\Lambda(t)$, over the duration of exposure years in the training data. From this vector of cumulative hazards, we derive a standardized life score that can be used to rank individuals for underwriting. Specifically, we select $\Lambda(10)$, the cumulative hazard at $t = 10$, corresponding to the median exposure of our data. The life score has a range of 0 to 100, ranging from highest to lowest risk. The score reflects the relative risk among five-year age-band, sex, and smoking-status cohorts — primary factors in actuarial mortality studies. Conditioned on cohort, the life score is the integer-valued percentile of the empirical distribution of all 10-year cumulative hazard values. Figure 3(a) shows that the proportion of each cohort is represented consistently across the range of life scores.

We can also demonstrate how medical impairments are stratified across the life score. Figures 3(b) and 3(c) display the proportion of heart condition incidence and BMI bands within each score decile. This highlights the effect that BMI and heart condition have on mortality risk. Each variable exhibits different stratification structures depending on its mortality dependence (for example, *U*- or *J*-shaped mortality curves; Chokshi, El-Sayed, and Stine [2015] and Cox et al. [2008]).

Model Validation

Traditional metrics that characterize predictive ability of statistical models, such as the concordance index for survival models (Harrell et al., 1982) or the area under the receiver-operating-characteristic curve commonly used in classification, are useful for research in model development, but they are insufficient to support business transformation and wide-scale adoption of a new paradigm for risk selection. In collaboration with business partners, we assess model performance and behavior across three levels — aggregate mortality impact; internal model smoothness and explainability; and predictions on individual cases.

Aggregate Mortality Impact

Actuaries focus on financial impact when developing and pricing life-insurance products through cash flow simulations that necessitate assumptions about expected mortality rates. Requiring a direct mortality comparison between model decisions and historical underwriting decisions, we designed a novel algorithm that generates a synthetic, model-assigned book of business. In the context of life underwriting, decisions refer to assigning applicants to one of several risk classes. MassMutual uses ultra-preferred nontobacco (UPNT), select-preferred nontobacco (SPNT), and standard (NT) nontobacco, and select-preferred (SPT) and standard (T) tobacco, in order of increasing risk. Substandard nontobacco and tobacco classes exist for additional medical impairments, and a small fraction may be declined for various financial and medical reasons.

The algorithm ensures that the number of simulated offers for each issue-year, risk-class, five-year age-band, sex, and smoking-status cohort are identical to those offered historically. This effectively controls for all actuarial factors and is consistent with how the life score is normalized. Without controlling for these factors, the algorithm would disproportionately assign, for instance, young females to the best risk classes, as they present low mortality risk.

The steps to generate equitable offers for an applicant pool can be illustrated through an example (refer to Maier et al. [2019] for more algorithmic details). First, assume that we know the number of policies by risk class within each cohort that were offered historically by underwriters, and that we have a life score for each historical applicant. Then, consider a cohort of 35-year-old, nonsmoking females who applied during 2005. Within this cohort, assume 100 applications were submitted, and underwriters offered fifty UPNT, fifteen SPNT, thirty NT, and declined coverage for five cases. Order the 100 cases by life score, assigning the fifty applicants with the highest life score to UPNT, the next fifteen, thirty, and five to SPNT, NT, and decline, respectively. Each 35-year-old, nonsmoking female who applied in 2005 now has a model- and underwriter-assigned risk class.

This simulation enables the calculation of business-relevant metrics, including the difference in deaths and the actual-to-expected (*A/E*) ratio for model assignment compared with underwriter assignment. The Society of Actuaries publishes a series of tables that contain the aggregate-mortality experience of the insured population across many carriers, and these tables are typically used as an expected baseline because they reflect a much larger population than that of a single carrier. The most recent tables, published in 2015, compile data from over fifty life insurers and facet mortality rates by age, gender, duration, and smoking status. Actuaries compare their observed, company-specific mortality experience against expected mortality rates with the *A/E* ratio, computed by summing observed deaths divided by the accumulated hazard corresponding to each individual policy-year on record:

$$A/E = \frac{\sum \text{event indicator}}{\sum \text{accumulated hazard}}$$

In our setting, the observed deaths stem from either actual company experience (the underwriter assignments) or from the simulated model assignments. An *A/E* under 100 percent indicates that the actual mortality experience is better than expected.

We applied this procedure to all historical life-insurance applications submitted 2000 to 2016, amounting to roughly 850,000 applications and over 13,000 deaths. Recall that risk classes determine premiums based on expected mortality rates. At MassMutual, the UPNT class corresponds to the lowest mortality rate and premium; thus, an effective model must assign the lowest-risk individuals to UPNT to maintain profitability. The model should also stratify high-risk individuals into appropriate classes.

Figure 4 shows the cumulative percent difference in UPNT claims in this simulated book of business generated by the mortality model and random assignment. Underwriters are experts at risk selection — as demonstrated by the nearly fifty-percent increase in claims through random assignment — yet the mortality model would have formed a UPNT offer pool with nine percent fewer deaths after fifteen years. The results aggregated across all risk classes are qualitatively similar, but the best risk selection by the model occurs in UPNT.

To measure performance of the model with an actuarial lens, we perform an *A/E* analysis. Tables 1 and 2 display confusion matrices of *A/E* ratios for the risk classes formed by the model and underwriters. All *A/E*s are normalized by the marginal of the underwriter-assigned best risk class (UPNT and SPT, respectively) so that values can be interpreted relative to underwriter performance. The model consistently produces lower mortality rates in each risk class and is substantially higher in the <NT and <T pools. The joint *A/E* ratios indicate that the model effectively disperses mortality risk in desired directions throughout the risk classes. Combined with underwriter decisions, there is potential for improved risk selection. For example, where they agree on UPNT, the mortality risk is eighty-six percent of the marginal.

The mortality model leverages fewer data sources than underwriters, who review additional requirements such as prescription drug histories, motor vehicle records, and financial data. As such, these results are conservative. An algorithmic underwriting system

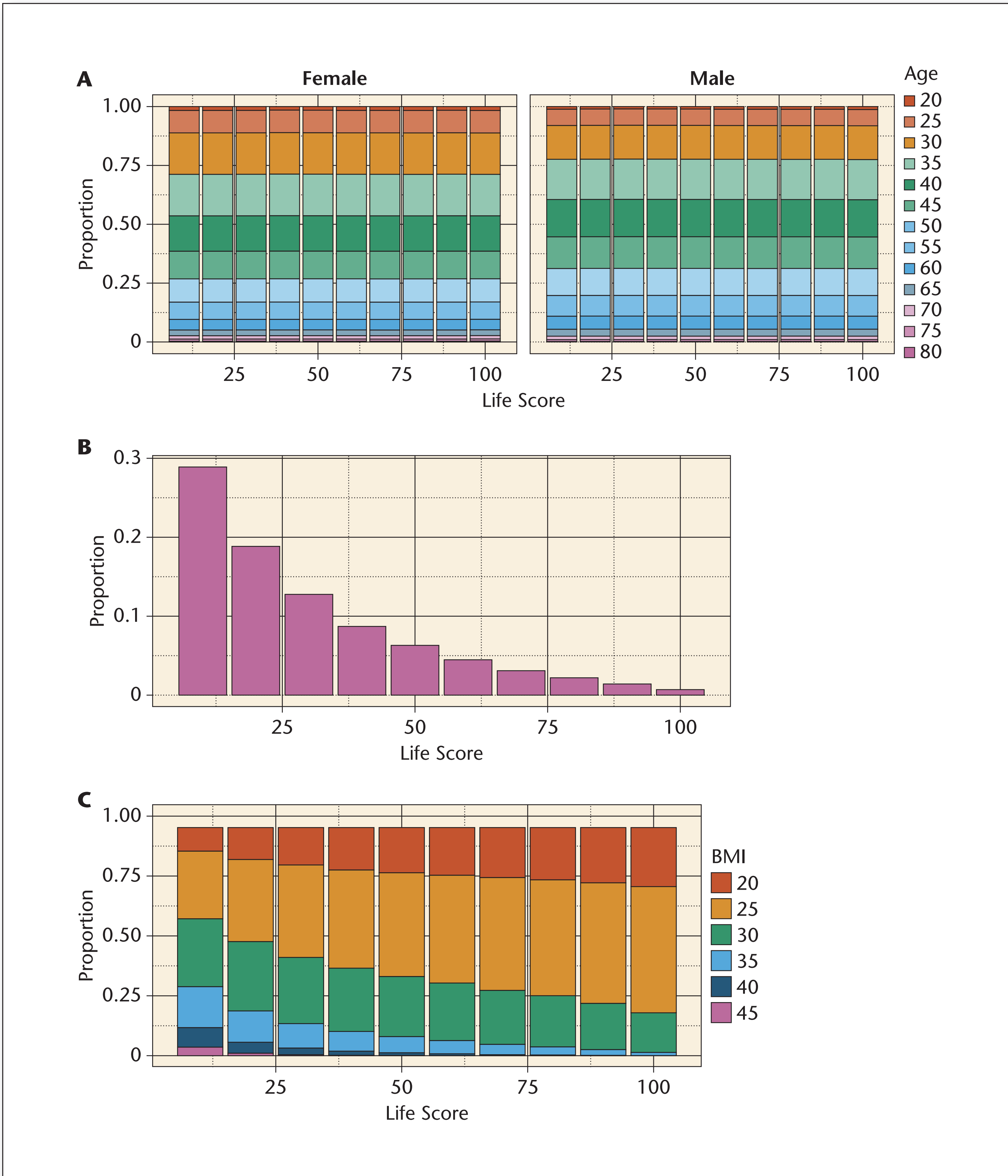

Figure 3. Performance of the Life Score.

(A) The proportion of individuals in each score decile is consistent across five-year age and sex bands. (B) Incidence of heart condition as a function of life score. The proportion ranges from twenty-nine percent in the first decile gradually decreasing to 0.9 percent in the tenth decile. (C) Distribution of BMI as a function of life score. The highest scores have a greater proportion of healthy-range BMI. As the score decreases, the proportion of the upper BMI extremes gradually increases.

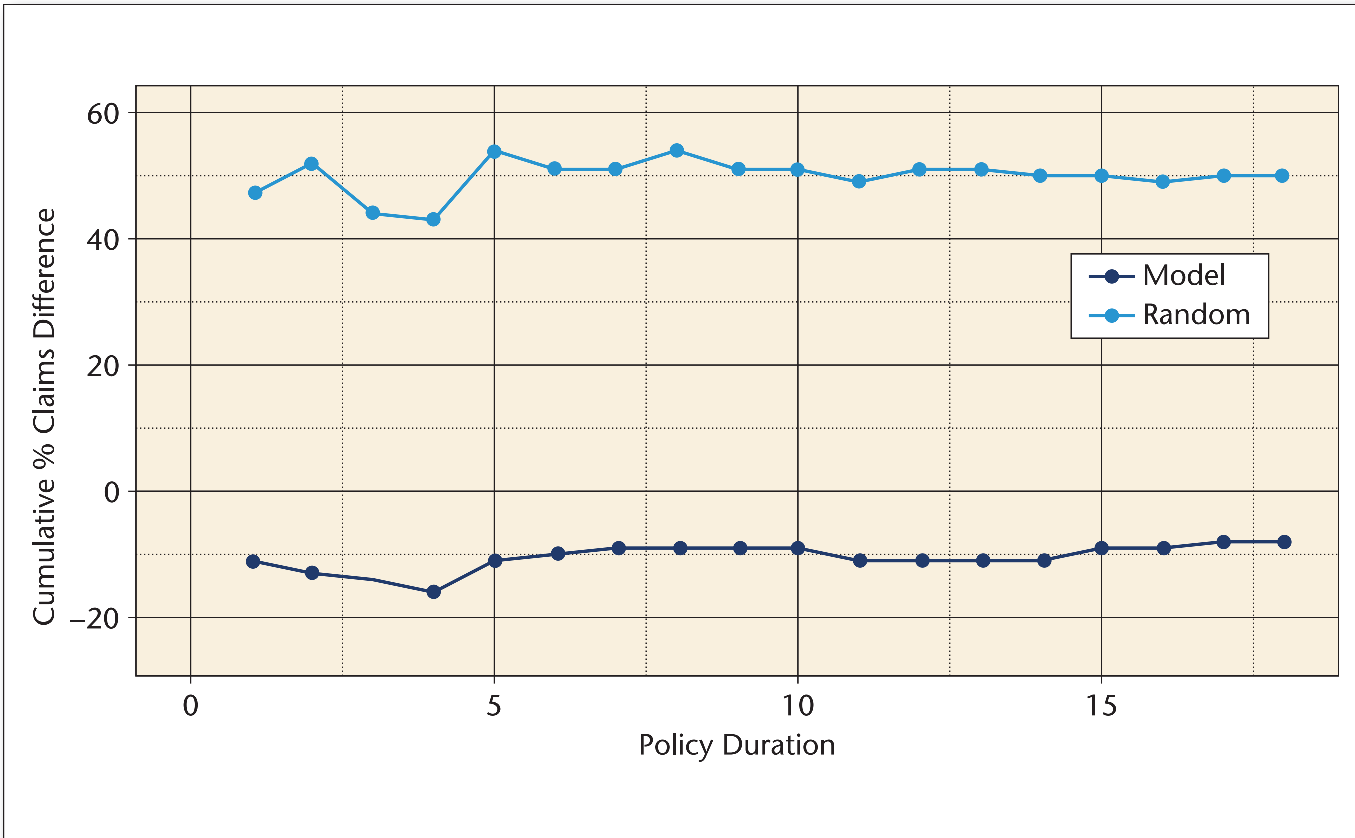

Figure 4. Cumulative Percent Difference in Deaths in UPNT Across Policy Duration, Where 0 Indicates Equivalent Counts.

Risk Class	UPNT	SPNT	NT	<NT	Marginal
UPNT	86	89	95	180	92
SPNT	97	113	137	177	117
NT	133	144	169	279	168
<NT	174	213	277	543	367
Marginal	100	119	160	363	

Rows: model; columns: underwriters.

Table 1. A/E Confusion Matrix for Nontobacco Classes Relative to UPNT.

Risk Class	SPT	T	<T	Marginal
SPT	71	78	102	76
T	122	122	178	131
<T	227	249	346	287
Marginal	100	126	235	

Rows: model; columns: underwriters.

Table 2. A/E Confusion Matrix for Tobacco Classes Relative to SPT.

combining the mortality model, a comprehensive rules environment, and controlled manual oversight can generate even stronger mortality results.

Internal Model Smoothness and Explainability

Understanding the inner workings of the model is crucial for ensuring it is robust and medically sound. Specifically, small fluctuations in input values should not lead to large changes in the resulting predictions, and the relative impact of each feature should align with medical expectations. Including these two conditions in the model validation process eliminates models that detect spurious correlations in the training set.

We assess internal model smoothness and explainability in two ways. First, box plots of the life score by grouped feature values, such as cholesterol in figure 5(a), ensure that, in aggregate, the life score is a smooth function of the given feature. The medical and data science teams review these plots to inspect the smoothness property of the mortality model. We expect median-centered life-score box plots at the healthiest range of inputs, such as normal total cholesterol.

The second approach uses SHAP values to assess the average marginal contribution of each input to the model. For example, figure 5(b) shows distributions of SHAP values for build-related features and blood proteins on a large set of scored cases. For many individuals, build and blood proteins have a small negative effect on overall model score, and for some individuals the effect is much higher. By comparing aggregate SHAP contributions, we can gauge relative feature importance and test how they align with medical knowledge.

Predictions on Individual Cases

Although some domains may only require strong overall model performance, life-insurance underwriting demands high-quality results at the individual level. If the model results do not match customer expectations of their own health, the company could lose business or be subject to antiselection.

As a final step of model validation, held-out cases are scored with the model for the medical team to review. The cases, typically numbering more than 150,000, are summarized and reviewed by the medical team to verify consistency with desired rates of offers within risk classes and to analyze trends in cases whose risk class would improve, worsen, or remain the same. A subset of several thousand recent applications are reviewed with more scrutiny, including model explanations that describe how each set of features contributed to the overall score. This pilot-review process ensures that the model performs well at an individual level.

Deployed Algorithmic Underwriting System

The validation of the mortality model provides sufficient evidence of its value, but it is a nontrivial undertaking to promote a model from a research environment to building a complete, real-time decision-making system.

The Algorithmic Underwriting System

A well-designed algorithmic underwriting system should capture digitally structured data and enable a simple interface and decision process for underwriters. Figure 6 depicts the high-level interactions among the major components of the algorithmic underwriting system at MassMutual. This includes the inputs and outputs for the mortality model, the rules engine that provides additional logic and coverage of underwriting guidelines, and the underwriter who serves as a human-in-the-loop to produce final underwriting decisions.

To begin the process, a prospective life-insurance customer completes a web-based application and, after a paramedic visit, has laboratory test results submitted on their behalf. To predict a life score, the mortality model requires inputs from these test results and responses within the health questionnaire portion of the application. Additional data, such as motor vehicle and prescription drug records, are obtained via vendor-supplied application-programming–interface calls after the applicant authorizes their disclosure. These underwriting requirements are necessary for complete underwriting, but are not included in the model due to limited historical coverage.

The same data are collected on applicants undergoing algorithmic and traditional underwriting, yet the processes are fundamentally different. Some of the technical and business challenges include generating discrete risk-class recommendations from continuous life scores; serving real-time scores in a robust environment; integrating model recommendations with medical and financial underwriting guidelines; and empowering underwriters with explanations behind individual life scores to enable communication with advisors and customers.

Calibrating Score Thresholds

The mortality model supports a flexible framework that can recommend risk classes based on different objectives. Because the life score measures mortality risk, actuaries could adjust offers to achieve desired levels of mortality. A simple approach sets thresholds that yield offer rates consistent with historical expectations and pricing assumptions.

Predicting in Real-Time

Once all requirements are received, the system sends a request to an internally developed application programming interface that hosts the mortality model. The technology behind this application programming interface is horizontally scalable, executes predictions in separate containers, and leverages established security and authorization protocols. The output of the model includes the predicted life score, recommended risk class, and explanations behind the prediction (see below). This response is transmitted within seconds, where the latency is driven by the complexity of the model prediction and derivation of explanations.

Integration With Underwriting Guidelines

Thousands of automated rules encompassing health, behavioral, and financial attributes serve as guardrails for the model. The rules reflect a comprehensive set of medical and underwriting guidelines developed and revised by experts in underwriting and insurance medicine. Each rule determines the best available risk class in the presence of certain values in the application. For example, certain medications may preclude an applicant from receiving a preferred offer. When a rule is triggered, underwriters can focus on pertinent details and use domain expertise to override the rule, allowing the case to continue through automated processing; decide if further investigation is warranted; or confirm the rule and proceed with the suggested risk class. The mortality model provides the basis for offers, but rules and professional review can lead to different ratings. This

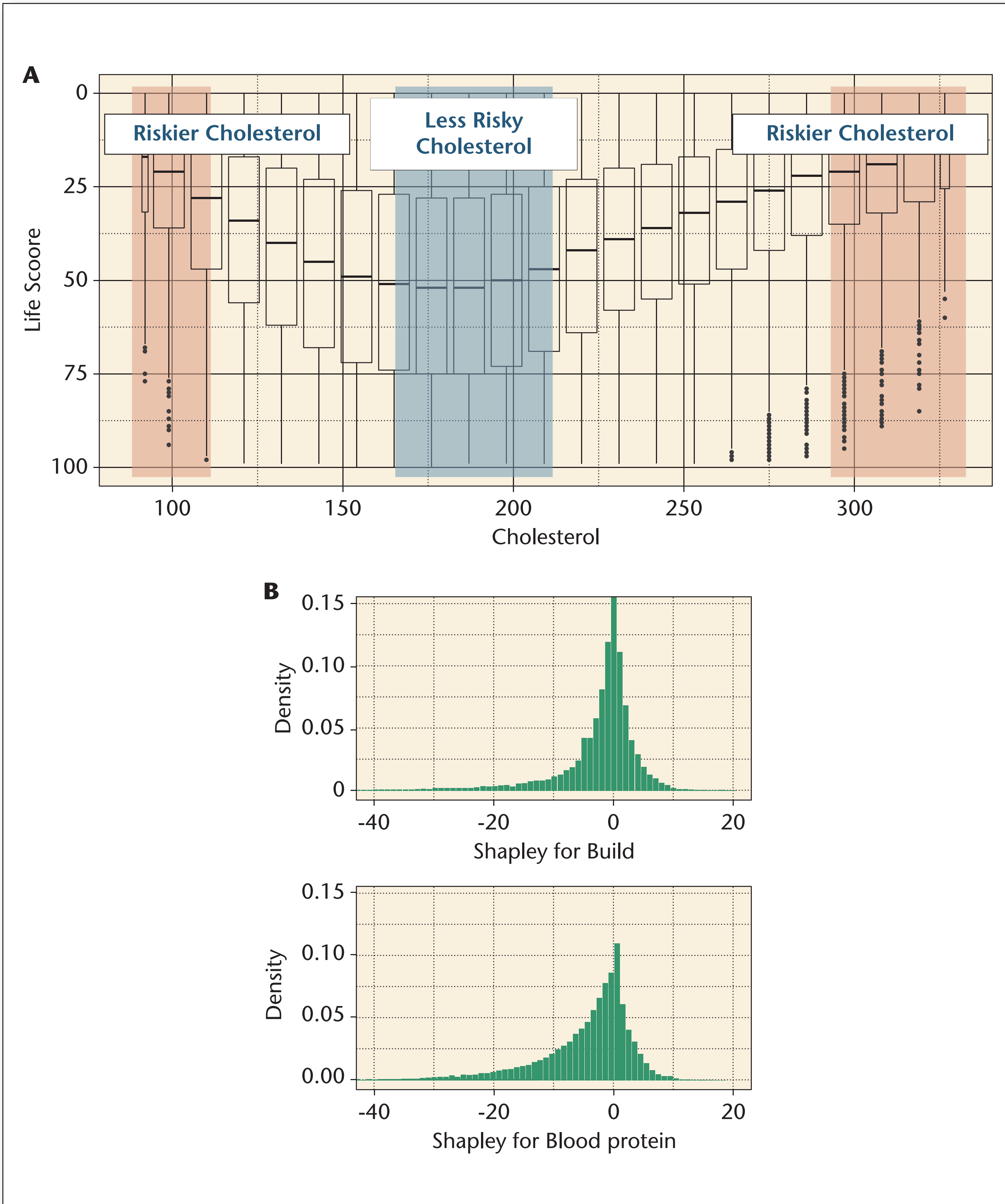

Figure 5. Assessing Smoothness of the Life Score.

(A) Plot of the life score showing a smoothly varying function of cholesterol, across regions of higher and lower risk. (B) SHAP contributions for build features (top) and blood protein features (bottom).

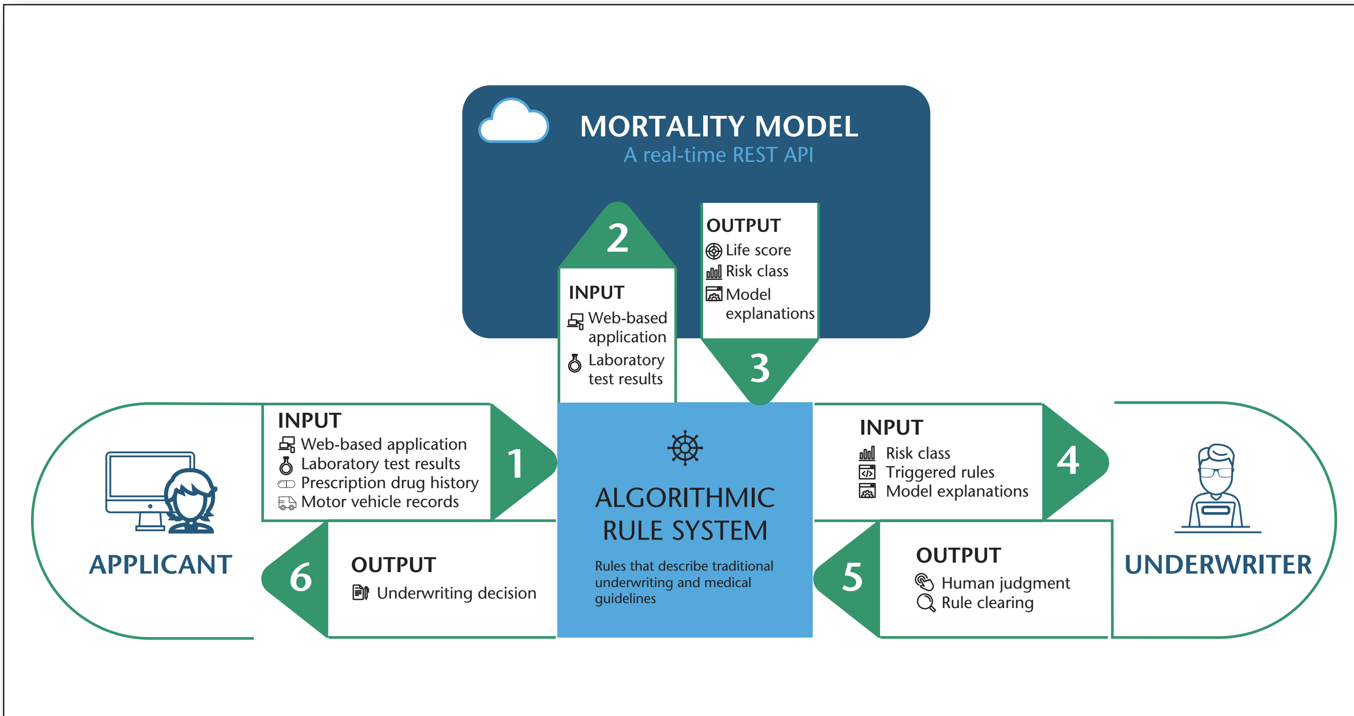

Figure 6. High-Level Schematic of the Algorithmic Underwriting System.

A life-insurance applicant submits several sources of underwriting data requirements. Laboratory test results and answers to the application are sent to the mortality model, which returns the life score, recommended risk class, and model explanations. The rules engine combines all underwriting data with the model output and presents the information to a human-in-the-loop underwriter who provides the final underwriting decision.

overall approach has led to new analyst positions and revised workflows for underwriters.

Interpreting Model Predictions

Model interpretability is an active area of research as machine-learning models become increasingly opaque. With a complex model driving risk-class decisions, it is imperative that underwriters can effectively explain why an applicant received a given offer. Recent research into SHAP values has led to a model-agnostic, theoretically justified, and computationally efficient framework (Lundberg and Lee, 2017). We adapt this methodology for the mortality model and compute feature contributions that compose individual predicted life scores. This provides sufficient explanatory information that can be displayed to underwriters handling each application.

Rolling Out the System

As a collaborative effort across various MassMutual teams, we systematically and gradually transitioned the exclusively human process of underwriting to an algorithmic framework. We conducted an initial pilot of the system in 2016 on 1,000 cases alongside traditional underwriting to compare risk-class assignments on new applications. After a successful pilot, algorithmic underwriting began issuing UPNT offers on term and whole life products up to $1 million benefits for applicants aged 17 to 40 years, followed by an expansion to $3 million and applicant ages up to 59 years, and finally for all standard-and-above risk classes. These product parameters cover approximately ninety percent of applications, and at present, the model has scored over 250,000 applications. In parallel, we integrated with the algorithmic underwriting platform developed by Haven Life[4] — a digital-first, direct-to-consumer life-insurance agency backed by MassMutual — and to date has scored nearly 20,000 applicants.

The implementation of predictive modeling in life underwriting has favorable implications for profitability and customer experience. At MassMutual, this has resulted in greatly improved operational efficiency — time to policy issuance has decreased by twenty percent for applicants in the healthiest risk class. This improvement has had material impact on customer experience as indicated by a twenty-five percent decrease of applicants declining to purchase their policies when the decision was made by the model compared with traditional underwriting within the best class. The automation of underwriting decisions at the company has amounted to labor savings of millions of dollars on a growing portfolio of policies with over one hundred billion dollars of protective benefits. Further, more profitability can be derived from the increased accuracy of the decisions when driven by the model; that is, the retrospective study of the mortality model suggests a long-term benefit of reduced claims experience. For these reasons, along with the additional drivers behind creating a standard life score, we also provide the life score as a service, through the establishment of LifeScore Labs, for use with other carriers that can benefit from improved underwriting.[5]

System Maintenance

Several teams support the monitoring, refreshing, and updating of the mortality model. Distributional drift, such as deteriorating offer rates, or sudden outliers, such as laboratory test changes, could affect the quality of decisions. We implement a monitoring protocol that reports on daily requests to the model and use web-based dashboards to visualize and track trends across time. An automated scoring process detects distributional differences in model inputs and outputs, and statistical anomalies trigger email alerts to model stewards.

The model is retrained annually to incorporate refreshed data and performance enhancements. Updates include refreshed death information and additionally underwritten applications. Collaborating with medical experts, enhancements to the model address concerns identified from case reviews. New versions have improved individual risk-class recommendations and transparency.

Prior to deploying new versions, we conduct retroactive pilots to avoid unexpected outcomes. Data scientists rescore recent cases and report aggregate statistics for the medical team to review prior to approval. Any change to the expected distribution of offers necessitates analysis and approval from actuaries. Final deployment requires collaboration among data scientists, data engineers who maintain the model application programming interface, and developers responsible for the production underwriting system.

Building Trust through Transparency

Customers are increasingly demanding improved experience and options for purchasing products. The life-insurance industry is responding with programs that leverage predictive models and algorithmic underwriting. We believe that proactive consideration of transparency is necessary for the success of the industry's digital transformation.

The Need for Increased Transparency

Increased accuracy and operational efficiency introduced by the mortality model and the algorithmic underwriting system are valuable improvements for pricing and scaling business. Conversely, the use of predictive modeling in life-insurance underwriting raises serious questions by consumer advocates and regulatory bodies, especially in light of scenarios for which algorithmic decisions have been questionable (for example, criminal sentencing [Angwin et al., 2016] and facial recognition [Raji et al., 2020]) and emerging regulation, such as the General Data Protection Regulation 2016/679 in Europe (Goodman and Flaxman, 2017), which require explanations behind algorithmic decisions. For life insurance, a simple legal mechanism — the Adverse Underwriting Decision letter — informs customers of the reasons behind certain pricing or coverage decisions, but this process has limited benefits for the industry.

The industry has an opportunity to proactively build consumer trust and meet current and expected future regulation by promoting transparency with prospective customers, applicants, underwriters, and regulators. A transparency-enabled life score can support education of how underwriting works for future customers and wide-scale adoption of consistent underwriting in the industry. Explaining how risk factors are considered in scoring individual applicants can empower underwriters to confidently interact with predictive models and provides helpful information for customers deciding whether to purchase offered coverage. And exposing the inner workings of models to regulators can provide assurance of compliance within specified legal frameworks.

A related issue to transparency, and of growing interest in consumer protection and regulation, is the concept of fairness and the impact that predictive models have on protected classes, such as race an ethnicity. The availability of a wide range of data sources that characterize individuals make life underwriting models and algorithmic decision-making vulnerable to persisting societal biases. However, the model described in this paper relies only on traditional underwriting requirements based on medical information with well-established causal ties to mortality risk. In addition, the life score is conditioned on age and sex such that applicants are compared relative to their cohort. Membership in other protected classes is intentionally unobserved during the underwriting process across the industry. Purposeful omission of these and other variables, such as geography and education, only partially mitigates concerns around fairness that stem from the use of algorithms in life underwriting. Defining, measuring, and adjusting for fairness remains an open challenge — just as it is also an active area of research in the machine-learning community (Holstein et al., 2019).

Implementing a Tool for Transparency

Statistical and machine-learning models have steadily improved in performance over the past few decades but, at the same time, the availability of tools for diagnosing and explaining model behavior has progressed more slowly, owing to both an underdefined objective and historically less emphasis by machine-learning researchers (Lipton, 2018). Fortunately, several recently proposed approaches provide

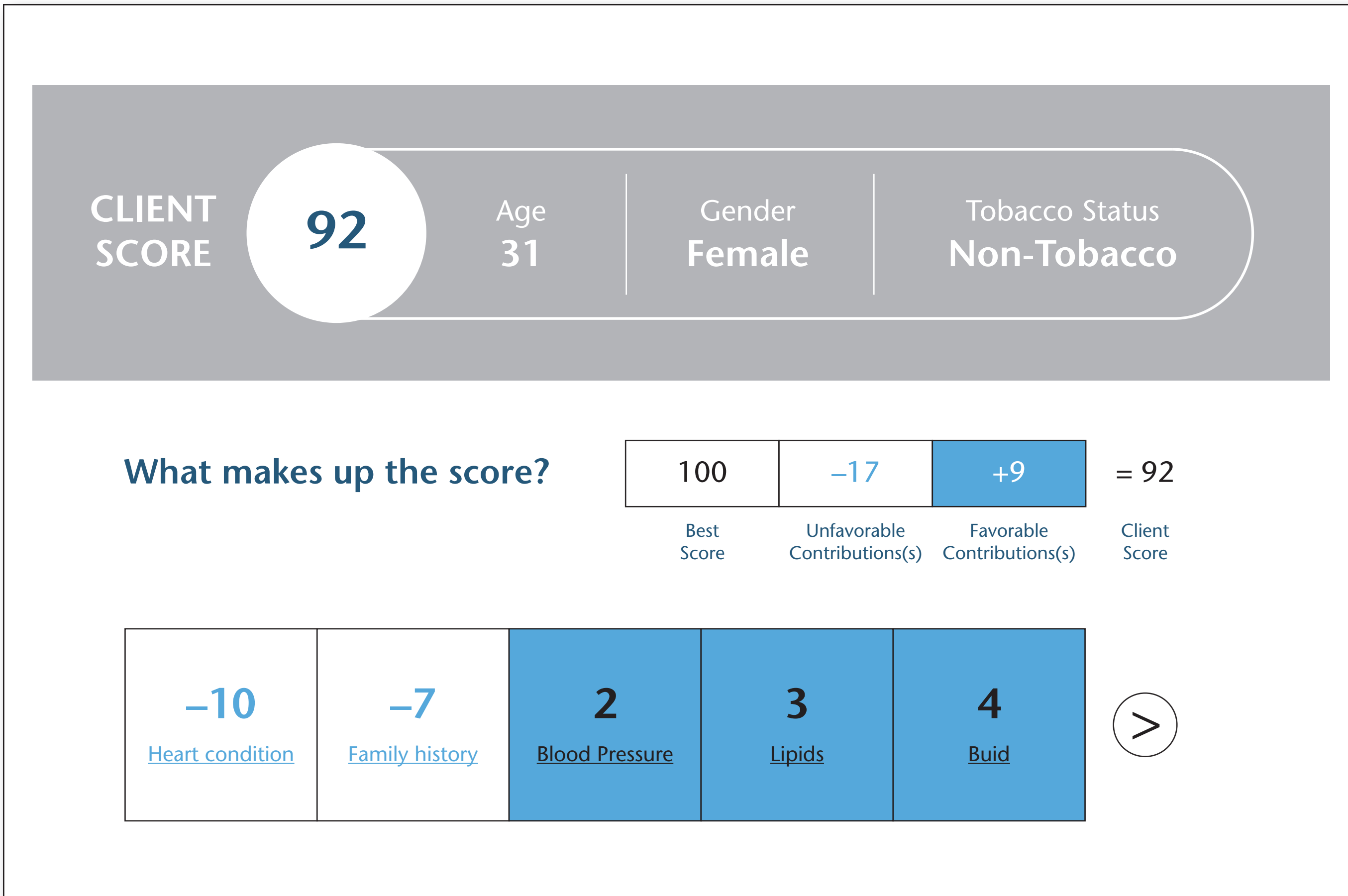

Figure 7. Transparency Displayed with MyLifeScore.

Feature contributions with MyLifeScore for a 31-year-old, nonsmoking female with a life score of 92. Unfavorable contributions subtract a total of seventeen points due to her heart condition and family history while favorable contributions related to her blood pressure, lipids, and build add nine points.

the transparency necessary for the successful adoption of predictive models in the life-insurance industry.

A family of model-agnostic, additive feature attribution methods aims to approximate model predictions as the sum of contributions made by model inputs, or features. These quantitative contributions can serve as an explanation, which can then be shared with individuals who interact with single-model decisions or aggregated across many predictions to provide details on population-level model behavior. These methods may not directly generate actionable explanations (Kumar et al., 2020), but when paired with appropriate domain knowledge, they can produce quantitative comparisons and insights behind individual predictions.

The SHAP framework developed by Lundberg and Lee, (2017) unifies other commonly used methods in this family, such as Local Interpretable Model-Agnostic "Explanations (Ribeiro, Singh, and Guestrin, 2016), Quantitative Input Influence (Datta, Sen, and Zick, 2016), and Deep Learning Important FeaTures (Shrikumar, Greenside, and Kundaje, 2017). SHAP is also model-agnostic, not requiring a differentiable model unlike the integrated gradients approach proposed by Sundararajan, Taly, and Yan (2017). Drawing from intuitive concepts in coalitional game theory, SHAP values have been shown to be the only additive feature attributions that are consistent and locally accurate. We implement a computationally efficient version of SHAP — the Kernel SHAP algorithm — to generate additive contributions of health attributes for each predicted life score. The contributions for an individual are compared relative to a baseline, healthy profile in the same five-year age, sex, and smoking cohort.

Feature contributions are returned by the mortality model in real-time during underwriting, as shown in figure 6. We define feature groupings by medically related categories, such as lipids and family history. To additionally facilitate education around how life underwriting functions and how different factors can drive insurance pricing, we developed a consumer-facing tool, called MyLifeScore, leveraging the same explanation framework of the mortality model. Figure 7 demonstrates how feature contributions can be displayed to consumers. Through greater visibility into the life score, we can provide consumers with more context and better expectations to make informed decisions in purchasing life insurance while also improving trust between consumers and carriers.

Conclusion and Future Directions

The emergence of large historical data sets and advancements in machine learning present an opportunity to improve the accuracy and transparency of underwriting in the life-insurance industry with standard measures of mortality risk. Leveraging 20 years of applications at MassMutual, we developed a mortality model and life score that can consistently compare applicants relative to their demographic cohorts. We demonstrated that embedding such an approach has profound implications for profitability and customer experience.

Avenues for future directions span data, methods, and insurance innovation. The current mortality model does not consider all traditional underwriting data sources, such as prescription drugs or motor vehicle records, and there are many nontraditional and unexplored sources, such as financial data, public records, electronic health data, wearable sensors, and genetic information, which may improve accuracy, enable alternative underwriting mechanisms, or enhance wellness programs to incentivize healthy behavior. Mortality models that do not rely on laboratory tests are of particular interest to many carriers seeking to launch or expand accelerated underwriting programs. Finally, machine-learning research on survival models may improve risk selection, and advancements in algorithmic fairness and transparency are equally crucial to study and implement.

Broad adoption of a standard measure of mortality risk opens the potential for exciting directions in the industry. We offered a few possibilities — increased access to insurance via consumer-driven purchasing, life-insurance–backed securities for portfolio diversification, and actionable wellness programs through quantifiable mortality benefits — and more industry-wide innovation will undoubtedly arise.

Acknowledgments

The authors are grateful for contributions made by Paul Shearer, Martha Grace, Ada Xu, Lizzie Kumar, John Karlen, and Debora Sujono. We also give thanks to our many colleagues at MassMutual, Haven Life, and LifeScore Labs, and our external collaborators for their continued partnership.

Notes

1. www.acli.com/Industry-Facts/Life-Insurers-Fact-Book
2. The MassMutual Mortality Score (M3S) and LifeScore Med360 refer to branded versions of the mortality model described in this work.
3. The consumer-facing MyLifeScore tool is openly available at lifescorelabs.com/products/mylifescore.
4. havenlife.com
5. lifescorelabs.com

References

Abrokwah, S.; Carroll, J.; Habecker, S.; Holzheu, T.; and Raturi, M. 2018. Life Underinsurance in the US: Bridging the USD 25 Trillion Mortality Protection Gap. Swiss Re Institute, September 1–18. www.swissre.com/dam/jcr:e8ea66fe-cc60-426f-8562-9fafbb4b4d83/expertise_publication_life_underinsurance.pdf.

Aggour, K. S.; Bonissone, P. P.; Cheetham, W. E.; and Messmer, R. P. 2006. Automating the Underwriting of Insurance Applications. *AI Magazine* 27(3): 36.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica (May 23). www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bender, A.; Groll, A.; and Scheipl, F. 2018. A Generalized Additive Model Approach to Time-to-Event Analysis. *Statistical Modelling* 18(3–4): 299–321. doi.org/10.1177/1471082X17748083.

Boodhun, N., and Jayabalan, M. 2018. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. *Complex and Intelligent Systems* 4(2): 145–54. doi.org/10.1007/s40747-018-0072-1.

Brackenridge, R. D. C.; Croxson, R.; and Mackenzie, R. 2006. *Brackenridge's Medical Selection of Life Risks*. Berlin, Germany: Springer.

Breiman, L. 2001. Random Forests. *Machine Learning* 45(1): 5–32. doi.org/10.1023/A:1010933404324.

Case, A., and Deaton, A. 2015. Rising Morbidity and Mortality in Midlife Among White Non-Hispanic Americans in the 21st Century. *Proceedings of the National Academy of Sciences of the United States of America* 112(49): 15078–83. doi.org/10.1073/pnas.1518393112.

Chokshi, D. A.; El-Sayed, A. M.; and Stine, N. W. 2015. J-Shaped Curves and Public Health. *Journal of the American Medical Association* 314(13): 1339–40. doi.org/10.1001/jama.2015.9566.

Cox, D. R. 1972. Regression Models and Life-Tables Regression. *Journal of the Royal Statistical Society. Series B. Methodological* 34: 187–220.

Cox, H. J.; Bhandari, S.; Rigby, A. S.; and Kilpatrick, E. S. 2008. Mortality at Low and High Estimated Glomerular Filtration Rate Values: A 'U' Shaped Curve. *Nephron. Clinical Practice* 110(2): c67–72. doi.org/10.1159/000151720.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *37th Institute of Electrical and Electronics Engineers (IEEE) Symposium on Security and Privacy*, 598–617. Los Alamitos, CA: Institute of Electrical and Electronics Engineers (IEEE) Computer Society. doi.org/10.1109/SP.2016.42.

Dubey, A.; Parida, T.; Birajdar, A.; Prajapati, A. K.; and Rane, S. 2018. Smart Underwriting System: An Intelligent Decision Support System For Insurance Approval & Risk Assessment. In *Proceedings of the Third International Conference for Convergence in Technology*, 1–6. Tamil Nadu, India: AIRCC Publishing Corporation.

Goldwasser, P., and Feldman, J. 1997. Association of Serum Albumin and Mortality Risk. *Journal of Clinical Epidemiology* 50(6): 693–703. doi.org/10.1016/S0895-4356(97)00015-2.

Goodman, B., and Flaxman, S. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine* 38(3): 50–7. doi.org/10.1609/aimag.v38i3.2741.

Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating The Yield of Medical Tests. *Journal of the American Medical Association* 247(18): 2543–6. doi.org/10.1001/jama.1982.03320430047030.

Holstein, K.; Wortman Vaughan, J.; Daume´, H. III; Dudik, M.; and Wallach, H. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 1–16. New York: Association for Computing Machinery (ACM). doi.org/10.1145/3290605.3300830.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random Survival Forests. *The Annals of Applied Statistics* 2(3): 841–60. doi.org/10.1214/08-AOAS169.

Katzman, J.; Shaham, U.; Bates, J.; Cloninger, A.; Jiang, T.; and Kluger, Y. 2016. *Deep Survival: A Deep Cox Proportional Hazards Network*. arXiv:1606.00931. Ithaca, NY: Cornell University Library.

Kumar, I. E.; Venkatasubramanian, S.; Scheidegger, C.; and Friedler, S. 2020. *Problems with Shapley-Value-Based Explanations as Feature Importance Measures*. Proceedings of the Thirty-Seventh International Conference on Machine Learning.

Lipton, Z. C. 2018. The Mythos of Model Interpretability. *Queue* 16(3): 31–57.

Lundberg, S. M., and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *In Advances in Neural Information Processing Systems*. 4765–74. La Jolla, CA: Neural Information Processing Systems Foundation, Inc.

Maier, M.; Carlotto, H.; Sanchez, F.; Balogun, S.; and Merritt, S. 2019. Transforming Underwriting in the Life Insurance Industry. In *Proceedings of the Thirty-First Association for the Advancement of Artificial Intelligence (AAAI) Conference on Innovative Applications of Artificial Intelligence*, 9373–80. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI) Press. doi.org/10.1609/aaai.v33i01.33019373.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)/Association for Computing Machinery (ACM) Conference on Artificial Intelligence (AI), Ethics, and Society*. New York: Association for Computing Machinery (ACM).

Ranganath, R.; Perotte, A.; Elhadad, N.; and Blei, D. 2016. *Deep Survival Analysis*. arXiv:1608.02158. Ithaca, NY: Cornell University Library.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the Twenty-Second Association for Computing Machinery (ACM) Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) International Conference on Knowledge Discovery and Data Mining*, 1135–44. New York: Association for Computing Machine (ACM). doi.org/10.1145/2939672.2939778.

Rosinger, A.; Carroll, M. D.; Lacher, D.; and Ogden, C. 2017. Trends in Total Cholesterol, Triglycerides, and Low-Density Lipoprotein In US Adults, 1999–2014. *JAMA Cardiology* 2(3): 339–41. doi.org/10.1001/jamacardio.2016.4396.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the Thirty-Fourth International Conference on Machine Learning*, 3145–53. Princeton, NJ: International Machine Learning Society, Inc.

Stout, R. L.; Fulks, M.; Dolan, V. F.; Magee, M. E.; and Suarez, L. 2007. Relationship of Hemoglobin a1c to Mortality in Nonsmoking Insurance Applicants. *Journal of Insurance Medicine (New York, N.Y.)* 39(3): 174.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Proceedings of the Thirty-Fourth International Conference on Machine Learning, 3319–28. proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf.

Wright, M. N., and Ziegler, A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77(1): 1–17. doi.org/10.18637/jss.v077.i01.

Marc Maier leads a data science team at MassMutual focused on improving underwriting, modernizing actuarial functions, and tackling societal issues around wellness, fairness, and transparency. He also previously launched and directed the Data Science Development Program at MassMutual. Maier earned a PhD in Computer Science from the University of Massachusetts Amherst.

Hayley Carlotto is a lead data scientist at MassMutual where she focuses on developing and deploying predictive models for underwriting programs at MassMutual and broadly in the life-insurance industry. Carlotto was featured in ThinkAdvisor's "30 Under 30" list in 2017 and completed MassMutual's Data Science Development Program in 2018. She holds an MS in Computer Science and a BS in Mathematics from the University of Massachusetts Amherst.

Sara Saperstein is a lead data scientist at MassMutual with over 10 years of AI experience in life insurance, advertising technology, and computational neuroscience. Her interests include survival modeling, predictions on continuously recorded signals, graph theoretic network analysis, and dynamical systems modeling. She holds an MA in Cognitive and Neural Systems from Boston University.

Freddie Sanchez is a data scientist in the Customer Journey group at MassMutual. His expertise lies in applying machine-learning techniques and experimentation to customer outreach campaigns. Sanchez earned an MS in Statistics from the University of Massachusetts Amherst.

Sherriff Balogun serves as chief of staff for MassMutual's Technology and Experience organization where he is responsible for ensuring strategic alignment, leading long-term strategic technology initiatives, and advising the chief information officer. Balogun previously held roles in data analytics at Springfield Public Schools and MassMutual. Balogun has an MS in Economics from Depaul University and a JD from the University of Connecticut Law School.

Sears Merritt leads the data, strategy, and architecture organization at MassMutual. Over the past 15 years, Merritt has spent time leading and innovating in numerous industries, including healthcare, telecommunications, and financial services. Merritt holds a PhD in Computer Science and an MS in Telecommunications from the University of Colorado Boulder and an MBA from the Sloan School at the Massachusetts Institute of Technology.