Transforming Underwriting in the Life Insurance Industry

Marc Maier, Hayley Carlotto, Freddie Sanchez, Sherriff Balogun, Sears Merritt
MassMutual Data Science
59 E Pleasant St, Amherst, Massachusetts 01002
{mmaier, hcarlotto, freddiesanchez, sbalogun, smerritt}@massmutual.com

Abstract

Life insurance provides trillions of dollars of financial security for hundreds of millions of individuals and families worldwide. Life insurance companies must accurately assess individual-level mortality risk to simultaneously maintain financial strength and price their products competitively. The traditional underwriting process used to assess this risk is based on manually examining an applicant's health, behavioral, and financial profile. The existence of large historical data sets provides an unprecedented opportunity for artificial intelligence and machine learning to transform underwriting in the life insurance industry. We present an overview of how a rich application data set and survival modeling were combined to develop a life score that has been deployed in an algorithmic underwriting system at MassMutual, an American mutual life insurance company serving millions of clients. Through a novel evaluation framework, we show that the life score outperforms traditional underwriting by 6% on the basis of claims. We describe how engagement with actuaries, medical doctors, underwriters, and reinsurers was paramount to building an algorithmic underwriting system with a predictive model at its core. Finally, we provide details of the deployed system and highlight its value, which includes saving millions of dollars in operational efficiency while driving the decisions behind tens of billions of dollars of benefits.

1 Introduction

Life insurance is a critical protective financial tool for millions of households. In the United States, life insurance companies collectively manage trillions of dollars of benefits. While there are numerous types of insurance contracts, a common component is the estimation of individual-level mortality risk through the process of underwriting. Traditionally, this is performed manually using human judgment and point-based systems that consider risk factors independently. These methods are sufficient in industry but are coarse and subject to inconsistency. As a result, traditional underwriting limits the degree to which an insurer can estimate risk from data and offer efficiently priced products.

The availability of large historical data sets provides an opportunity for machine learning to transform underwriting for life insurance. MassMutual, a large insurance and financial services company, has curated a data set of nearly one million applicants spanning 15 years and containing health, behavioral, and financial attributes. To the best of our knowledge, this is the largest and most comprehensive application data set in the industry. Combining this data with advancements in machine learning and survival modeling enables accurate estimation of mortality risk. We develop a high-resolution model that generates a *life score* and underpins the MassMutual Mortality Score (M3S) and LifeScore360.[1]

Collaborating with actuaries, we design a novel evaluation framework to compare historical underwriting decisions against simulated model decisions over a 15-year period. This empirical study demonstrates that the life score outperforms traditional underwriting, yielding a 6% reduction in claims in the healthiest pool of applicants. Based on these promising results, we engaged additional partners across MassMutual to implement an algorithmic underwriting system with this mortality model as its primary risk-driving engine. Over the past two years, this system has reduced time to issue by $>25\%$ and increased customer acceptance by $>30\%$ for offers made with light manual review, while saving millions of dollars in operational efficiency and driving the decisions behind tens of billions of dollars of benefits.

The remainder of this paper: (1) provides background on life insurance and the mathematical frameworks used to quantify risk in insurance; (2) describes the data set and methodologies used to estimate mortality risk; (3) presents performance results and deployment details; and (4) discusses the future and implications of using predictive models as a core component of underwriting in life insurance.

2 Background

This section provides background on the traditional underwriting process, survival modeling, and actuarial science.

2.1 Life Insurance and Underwriting

A life insurance policy is an agreement between a policyholder and an insurer whereby the insurer agrees to pay beneficiaries a sum of money at the time of the policyholder's death. In return, the policyholder pays premiums over a predefined period of time (Atkinson and Dallas 2000). Life insurance provides security to the beneficiaries by reducing

[1] M3S and LifeScore360 refer to branded versions of the mortality model described in this work.

the financial impact of an untimely death. Beneficiaries can use the proceeds to pay for future expenses (e.g., daily living expenses, college tuition, retirement) that would have otherwise been paid for by the earnings of the insured.

Most types of life insurance require an estimate of 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 (Black and Skipper 2000). Actuaries compute the cost of covering mortality risk over the lifetime of the policy and translate it into a set of premium payments (Jordan 1967). The financial risk and general approval of the underwriting process is agreed upon with reinsurance companies, institutions that assume a portion of the risk and who diversify their holdings across insurance industries.

In contrast to other types of insurance, such as auto, home, and health, life insurance is typically purchased through a financial advisor who connects an individual to a carrier and helps clients identify the type and amount of insurance that suits their needs. Advisors provide estimates of the premiums, but the exact price is determined after underwriting.

Life insurance underwriting has primarily used point systems developed by doctors and underwriters. These systems calculate risk by mapping medical and behavioral attributes—such as cholesterol, build, driving record, and family and personal medical history—to point values that either debit or credit an overall score (Brackenridge, Croxson, and Mackenzie 2006). This approach resembles risk calculations employed in clinical medicine (e.g., Framingham risk scores (Wilson et al. 1998)). A life underwriter reviews an application to calculate the net number of points, determining one of several risk classes that drive premium and are priced according to aggregate mortality.[2] Advancements in statistics and machine learning present an opportunity to update the traditional approach to underwriting, which predominately considers factors independently. Leveraging AI to automate underwriting decisions is not novel in the industry (e.g., using fuzzy logic (Aggour et al. 2006)), but developing a machine learning model that outperforms human decisions and deploying at scale is unprecedented.

2.2 Survival Modeling

The majority of predictive modeling tasks are based on classification or regression. In the context of survival analysis, however, the outcome of interest is the duration until a binary event may occur for a particular record. The objective of survival analysis is to approximate the *survival function*, $S(t) = Pr(T > t)$, which describes the probability that an event, occurring at random variable time T, occurs later than some given time t. The *hazard rate*,

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)}, \tag{1}$$

is the rate of the event at time t conditioned on having survived until time t. In actuarial science, the hazard is often denoted as μ and describes the mortality rate for a given attained age. The *cumulative hazard function*, defined as

$$\Lambda(t) = \int_0^t \lambda(u)du, \tag{2}$$

is related to the survival function as $\Lambda(t) = -\log S(t)$. Nonparametric estimators, namely the Kaplan-Meier (Kaplan and Meier 1958) and Nelson-Aalen estimators, compute these quantities directly from observed survival data.

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. In the underwriting-for-mortality setting, the covariates are medical and behavioral attributes of life insurance applicants and the event is mortality. The techniques used to estimate these functions fundamentally require a different set of statistics as the time-to-event of mortality is unknown for most individuals. This is referred to as *right-censored* data because the date of birth is known, but the date of death is unobserved for a large set of individuals. Missing survival information is a key characteristic of survival analysis, in which the data may be censored at the beginning, end, or even middle of study periods.

There is a well-established set of methods employed by academic and industrial practitioners of survival analysis. The Cox proportional hazards model is the most widely used statistical technique for estimating individual risk in studies of survival (Cox 1972). This is a semi-parametric regression model that assumes a linear functional form and proportional hazards for any two strata over time. In machine learning, random forests (Breiman 2001) have been adapted by Ishwaran (2008) to handle right-censored survival outcomes (called random survival forests, or RSF) and efficient implementations exist (Wright and Ziegler 2017). As a nonparametric, adaptive model, RSF captures interactions and non-linear dependencies that are more subtle and complex than can be reflected by a linear model. The extension to survival data includes setting the splitting criterion to maximize survival difference, as measured by a log-rank test, and the terminal nodes directly estimate the cumulative hazard function via an ensemble of Nelson-Aalen estimators.

Survival models can be evaluated with *concordance*, a pairwise ranking statistic similar in interpretation to area under the receiver operating characteristic curve (AUC) commonly used in classification. The next section provides background on a more relevant metric for an actuarial setting.

2.3 Actuarial Mathematics

Actuaries evaluate mortality risk and its financial impact when developing life insurance products. Pricing and cash flow simulations require assumptions about expected mortality rates. These are derived from a combination of observed mortality experience within a company and industry-wide life tables. The Society of Actuaries publishes a series of *Valuation Basic Tables* (VBTs) that aggregate mortality experience within the insured population across many carriers. The most recent tables, published in 2015, compile data

[2] MassMutual uses the following risk classes: ultra-preferred (UPNT), select-preferred (SPNT), and standard (NT) non-tobacco and select-preferred (SPT) and standard (T) tobacco, in order of increasing risk. Substandard non-tobacco and tobacco classes exist for specific medical impairments, and a small fraction may be declined for various financial and medical reasons.

from over 50 life insurers and facet mortality rates by standard factors: age, gender, duration, and smoking status.

VBTs are often used as a standard baseline because they reflect a much larger population than that of a single carrier. Actuaries compare their observed mortality experience against the expected mortality rates in the VBTs using a metric referred to as the *actual-to-expected* (A/E) ratio. The A/E ratio is computed by summing all observed deaths divided by the accumulated hazard corresponding to each individual policy year on record:

$$A/E = \frac{\sum \text{event indicator}}{\sum \text{accumulated hazard}}. \tag{3}$$

When the A/E is less than 100%, this indicates that the actual mortality experience is better than expected. In this work, we rely on A/E ratios to compare model performance against underwriters using the 2015 VBT expected basis.

3 Modeling Mortality

With an understanding of traditional underwriting for life insurance, this section demonstrates how historical underwriting data can be leveraged to train models of mortality.

3.1 Data

Life insurance carriers track policies over a potentially long period of time to maintain records of financial exposure. Minimally, this requires demographics, policy details, and post-issue events (e.g., status changes). However, to build a model that predicts mortality risk, it is critical to retain data used for underwriting. MassMutual has a consolidated, digital record of nearly one million applications for which a lab test was ordered during 1999–2014. After removing applications with a high degree of missing values, typically incomplete or withdrawn applications, this reduces to 908k records with 9.16M exposure years and 15.7k observed deaths.

To develop a general-purpose model for life underwriting, mortality outcomes on all applicants are crucial. Internal records are limited to death benefit claims, which excludes applicants that never became policyholders or terminated their policies prior to a claim. MassMutual obtained and periodically refreshes ground-truth mortality data from internal and third-party sources on historical applicants.

Aside from demographics, labs, and mortality, the data set covers attributes drawn from a lengthy health history questionnaire that accompanies the application process. Additional data sources are widely used in underwriting, such as prescription drug histories and motor vehicle records, but at present, we do not have adequate historical coverage to directly tie to mortality. Below we review select statistics on the primary attributes contained in the overall data set.

Demographics Over a 15-year period, the data set provides broad coverage of demographics. Males are generally older than their female counterparts at time of application, as shown in Figure 1. Males account for more than twice the number of deaths, which is a function of an older age distribution, a higher proportion in earlier application years, and a higher mortality rate in general (see Figure 2). Additionally, the applicant data exhibit expected survival dependence on age and sex (e.g., females tend to outlive males (Kalben 2000) and survival probability decreases with age).

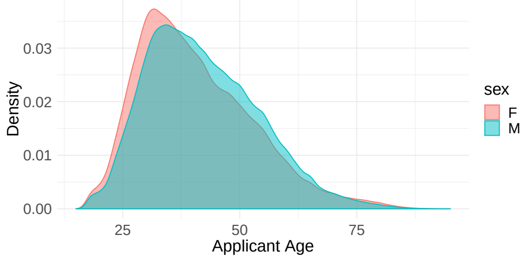

Figure 1: The age distribution of the application population stratified by sex.

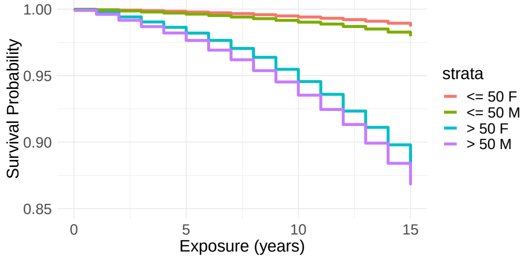

Figure 2: Survival probabilities by age $\leq$ 50 and sex are consistent with general population statistics.

Lab Tests Life insurance underwriting typically includes a *de facto* 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 (Goldwasser and Feldman 1997) and cholesterol (Kronmal et al. 1993). The lab test data provide broad exposure to a range of values and includes biophysical measurements (e.g., build, blood pressure), lipids (e.g., cholesterol), liver function tests (e.g., gamma-glutamyltransferase), kidney function tests (e.g., creatinine), blood proteins (e.g., albumin, globulin), urine proteins (e.g., microalbumin), blood sugars (e.g., fructosamine, hemoglobin A1C), and several indicators (e.g., cocaine, HIV).

Health History Questionnaires Lab tests are a point-in-time view into an individual's health that yield substantial protective value for risk selection. The application process also solicits information related to personal and family health history, as well as behavioral risk through an extensive questionnaire. 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. Despite the acquisition costs, this data enable a consistent mapping with the current application. The training data include variables that align to major medical impairments (e.g., cardiovascular disease) derived from Boolean responses and keyword extraction on open-text fields.

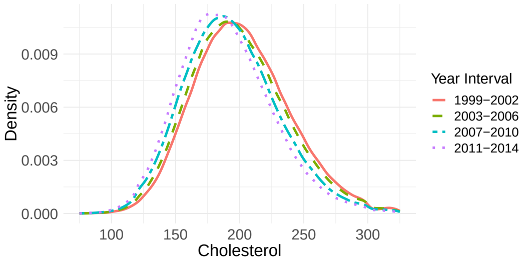

Figure 3: Grouping by 4-year bands, the distribution of cholesterol trends lower over time.

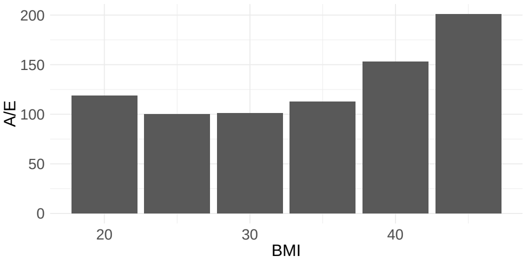

Figure 4: Trends in aggregate mortality risk, measured by A/E, as a function of 5-point bands of BMI.

Health Trends across Time Given the 15-year time period of our data, we observe trends in the distribution of certain lab values. For example, recent applicants exhibit lower levels of cholesterol compared to those in earlier years, as shown in Figure 3. This is consistent with medical research reporting similar trends over the same time period (Rosinger et al. 2017). A variable that trends over time is referred to as covariate shift or non-stationarity, which presents a modeling challenge due to the temporal association with predictive variables (Sugiyama and Kawanabe 2012). 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 lab tests and mortality risk.

3.2 Modeling

This valuable data asset enables survival modeling. Below, we outline the strategy for selecting relevant features and refining the mortality model, and we demonstrate that the predictions appropriately stratify health factors.

Feature Selection Feature selection was heavily influenced by medical and actuarial experts and validated with standard machine learning techniques. A model intended to be used for an embedded and central process to the business cannot solely be optimized for predictive accuracy. It is critical to consider the operational impact of each prediction, including reconciling with complementary underwriting data sources and transparent communication to customers.

Through close partnership with the MassMutual's medical team, we constructed an intuitive and medically relevant mortality model. Given their recommendations, we reviewed the historical coverage of each variable as procedures for filing and testing have changed across time and underwriting requirements vary by demographics and policy features. We also assessed the statistical dependence with mortality inherent to each variable. For example, Figure 4 shows how A/E ratios vary by 5-point bands of body mass index (BMI), exhibiting elevated mortality risk for low BMI and steadily increasing mortality risk for higher values of BMI.

The deployed mortality model relies on nearly sixty inputs, including internally generated features (e.g., BMI as a function of height and weight). The main inputs are captured in biophysical measurements, blood and urine specimens, and applicant health history questionnaires.

Experiments Research on survival methods has made advances over the past decade. The most widely studied machine learning models for survival data are tree-based methods, such as the random survival forest. Emerging research aims to apply advanced statistical models, such as gradient boosting and generalized additive models to discrete-time survival analysis (Chen and Guestrin 2016; Wood 2006), 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 environment. Thus, we focused our modeling on the Cox proportional hazards model (COX) and random survival forest (RSF). Experiments iterated on findings drawn from our collaborative feature selection process, in addition to improvements through variable transformation, hyperparameter tuning, and sampling techniques. Each experiment performed 10-fold cross validation and held-out predictions were used to produce a suite of statistical, actuarial, and business-relevant evaluation metrics. The RSF model consistently yields a substantial improvement over traditional underwriting and COX.

Developing the Life Score The RSF mortality model directly estimates the cumulative hazard function, $\Lambda(t)$, across the duration of exposure years in the training data. From this vector of cumulative hazards, we derive a single, 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 0–100, ranging from highest to lowest risk. The score reflects the relative risk among 5-year age band, sex, and smoker cohorts—primary factors in actuarial mortality studies. Conditioned on cohort, the life score is the integer-valued quantiles of the empirical distribution of all 10-year cumulative hazard values. Figure 5(a) demonstrates that, as expected, the proportion of each cohort is represented consistently across the range of life scores.
Example: If Carlos is a 55-year-old non-smoking male with a life score of 87, he can be compared directly against and has lower mortality risk than Barry, another 55-year-old non-smoking male with a score of 53. However, if Amy is a 35-year-old non-smoking female with a score of 87, she does not necessarily present the same mortality risk as Carlos.

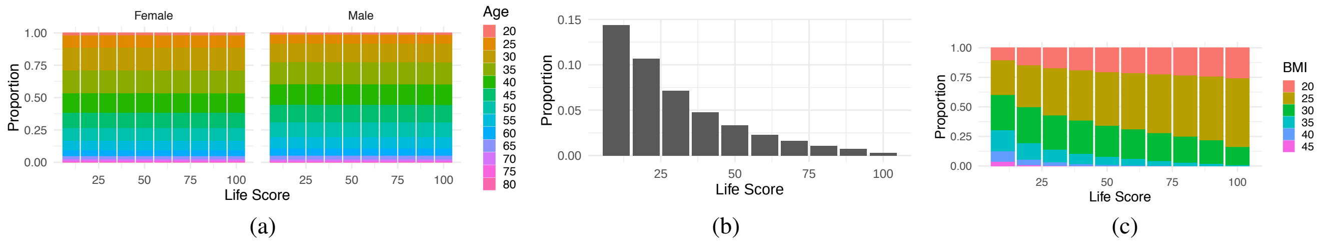

Figure 5: (a) The proportion of individuals in each decile of the score is consistent across 5-year age and sex bands. (b) Incidence of heart condition as a function of life score. The proportion ranges from 14.4% in the first decile to 0.3% in the tenth decile, gradually decreasing in between. (c) Distribution 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 upper and lower BMI extremes gradually increases.

We can also demonstrate how medical impairments are stratified across the life score. Figures 5(b) and 5(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 (e.g., U- or J-shaped mortality curves (Chokshi, El-Sayed, and Stine 2015; Cox et al. 2008)).

4 Validation

Analyses of correlations among the life score and health factors are useful, but business-related metrics are critical to understand the expected performance of a deployed system.

4.1 Simulation Method

In collaboration with actuaries, we designed a novel algorithm that generates a synthetic, model-assigned book of business to compare against historical underwriting risk class offers. The algorithm ensures that the number of simulated offers for each issue year, risk class, 5-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 example, young females to the best risk classes as they present lower mortality risk.

Algorithm 1: AssignRiskClasses(D, M)

```
$D_{assign} \leftarrow \emptyset$
for $Year\ y, Age\ a, Sex\ s, Smoking\ Status\ t$ do
    $D_{cohort} \leftarrow D[Y = y, A = a, S = s, T = t][:]$
    $offer\_counts \leftarrow count(D_{cohort})$
    $D_{cohort}[:][LS] \leftarrow predict(D_{cohort}, M)$
    $D_{cohort} \leftarrow sort(D_{cohort}[:][LS])$
    $idx \leftarrow 1$
    for $ordered\ Risk\ Class\ r$ do
        $D_{cohort}[idx : offer\_counts[r]][r_{model}] \leftarrow r$
        $idx += offer\_counts[r]$
    $D_{cohort}[idx :][r_{model}] \leftarrow decline$
    $D_{assign} \cup= D_{cohort}$
return $D_{assigned}$
```

Table 1: A/E confusion matrices for (a) non-tobacco classes relative to UPNT and (b) tobacco classes relative to SPT. (rows - model; columns - underwriters)

(a)

	UPNT	SPNT	NT	<NT	Marginal
UPNT	84	85	109	177	93
SPNT	100	120	143	256	127
NT	126	143	174	247	163
<NT	226	306	340	653	432
Marginal	100	126	174	381	

(b)

	SPT	T	<T	Marginal
SPT	68	79	156	80
T	107	148	149	130
<T	287	274	409	329
Marginal	100	142	253	

The steps to equitably generate historical offers for a pool of applicants is shown in Algorithm 1. Using the historical data D, the algorithm first computes the number of offered policies by risk class within each cohort. Then, the mortality model M predicts a life score LS for each individual in D_{cohort}. For each risk class r in order, assign r to the next $offer\_counts[r]$ lowest-risk individuals that have yet to be assigned a model risk class, r_{model}. Assign a worse-than-standard rating to the remaining individuals.
Example: Consider a cohort of 35-year-old, non-smoking females in 2005. Assume 100 applications were submitted, and underwriters offered 50 UPNT, 15 SPNT, 30 NT, and declined coverage for 5 cases. Order the cases by life score and assign the 50 applicants with the highest life score to UPNT, the next 15, 30, and 5 to SPNT, NT, and decline, respectively. Each 35-year-old, non-smoking female who applied in 2005 now has a model- and underwriter-assigned risk class.

4.2 Simulation Results

The model-assigned risk classes produced from Algorithm 1 enable the calculation of useful statistics, including the difference in deaths and A/E ratios compared to underwriters. We applied this simulation to historical life insurance ap-

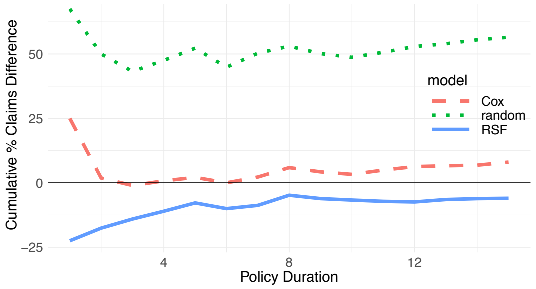

Figure 6: Cumulative percent difference in deaths in UPNT across policy duration, where 0 indicates equivalent counts.

plications submitted 2000–2014 and assume policies remain active until death, ignoring lapse. This amounts to roughly 650k applications and 7k deaths.

Recall from Section 2.1, risk classes determine premiums based on expected mortality rates. The UPNT class corresponds to the lowest mortality rate and premium; thus, an effective model must assign those lowest-risk individuals to UPNT to maintain profitability. The model should also stratify high-risk individuals into the appropriate classes.

Using the output of Algorithm 1, we compute the difference in death counts for the RSF mortality model, as well as COX and a random scoring process. Figure 6 displays the cumulative percent difference in UPNT deaths for the three methods compared to underwriters. Underwriters are experts at risk selection, yet the results show that after a 15-year duration, RSF would have formed an offer pool with 6% fewer deaths. COX and the random process produce 8% and 57% more deaths than underwriters, respectively. The results aggregated across all risk classes are qualitatively similar.

To measure performance of the RSF model with an actuarial lens, we perform an A/E analysis. Tables 1a and 1b display confusion matrices of A/E ratios for the risk classes formed by RSF and underwriters. All A/Es 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 RSF model consistently produces lower mortality rates in each risk class and is substantially higher in the <NT and <T pools. The joint A/Es 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 84% 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 combining the mortality model, a comprehensive rules environment, and controlled manual oversight will generate even better mortality results.

5 Deployment

The simulation study illustrates the value of the mortality model, but it is a non-trivial undertaking to promote a model from a research environment to a real-time decision-making system. Below, we describe the approach to developing, releasing, and monitoring an algorithmic underwriting system.

5.1 The Algorithmic Underwriting System

A well-designed algorithmic underwriting system should capture digitally structured data and enable a simple interface and decision process for underwriter interaction. At MassMutual, a prospective life insurance customer completes a digital application and submits laboratory tests, generally through a paramedic visit. To predict a life score, the mortality model requires inputs from these lab test results and responses within the health questionnaire portion of the application. Additional information required for underwriting, such as motor vehicle records and prescription drug history, is obtained via vendor-supplied API calls. This information is not included in the model as historical coverage of this data is currently limited. The same data are collected on applicants undergoing algorithmic and traditional underwriting, yet the overall processes are fundamentally different. Some of the technical and business challenges include (1) generating discrete risk class recommendations from continuous life scores; (2) serving real-time scores in a robust environment; (3) integrating the model recommendations with medical and financial underwriting guidelines; and (4) empowering underwriters with explanations of the factors 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. For example, because the life score measures mortality risk, actuaries could adjust offers to achieve desired levels of mortality. However, the current approach sets thresholds that yield offer rates consistent with historical rates as those form the basis of pricing assumptions. This aligns with the design of the simulation study from Section 4.1 and its corresponding metrics.

Predicting in real time. Real-time risk class recommendations are accessed via an internally developed REST API that hosts the mortality model. Once the full set of requirements are received for an application, the algorithmic underwriting system sends a formatted request to the API to receive the life score and suggested risk class. The API is highly scalable and responds within seconds, where the latency is driven by the complexity of the model prediction.

Integration with underwriting guidelines. Thousands of automated rules encompassing health, behavioral, and financial attributes serve as guardrails for the risk class recommendations generated by the model. The rules reflect a comprehensive set of medical and underwriting guidelines developed 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, a high BMI would preclude an applicant from receiving a preferred offer. When a rule is triggered, underwriters can focus on pertinent details of the application and use domain expertise to (1) override the rule, allowing the case to continue through the automated process, (2) decide if additional information is required for further review, or (3) confirm the rule and proceed with the suggested risk class. Ultimately, the life score drives the final offer, but the rules may lead to a worse rating. This approach to underwriting has led to new analyst positions and revised workflows for underwriters.

Interpreting model predictions. With a complex model driving risk class decisions, it is imperative that analysts and underwriters can effectively explain why an individual applicant received a given offer. Model interpretability is an active area of research as machine learning models become increasingly opaque, despite evidence that even linear models can present a challenge in its interpretation (Lipton 2016). We developed a model-agnostic approach to generating interpretable, approximate factors that contribute to the life score at an individual prediction level. The methodology is similar to recent research, including Shapley values and LIME (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016). The contribution factors are returned with the life score and displayed to underwriters.

5.2 Rolling out the System

We systematically and gradually transitioned the exclusively human process of underwriting to an algorithmic framework. As a proof-of-concept, we conducted a pilot of the system on 1,000 cases in parallel to traditional underwriting. This enabled observation of risk class offer rates and agreement between the two systems. Following a successful pilot, algorithmic underwriting began issuing UPNT offers on all life products up to \$1M benefits for applicants aged 17–40. This was followed by an expansion to \$3M and applicant ages up to 59, and finally for all standard-and-above risk classes. At present for these parameters, algorithmic underwriting is applied to 90% of applications.

5.3 System Maintenance

Collaboration across several teams supports the monitoring, refreshing, and updating of the mortality model. A critical component to an algorithmic underwriting system, or any machine learning system (Sculley et al. 2015), is to continuously monitor the model inputs and outputs. Distributional drift, such as deteriorating offer rates, or sudden outliers, such as a lab test changing units, could manifest in the system, affecting the quality of the decisions. We implement a monitoring protocol that reports on daily batches of requests to the mortality model and use web-based dashboards to visualize and track trends across time. In the future, we plan to establish automated monitoring to detect anomalies and structural changes to model inputs and risk class offer rates.

The model is retrained periodically to incorporate refreshed data and performance enhancements. Updates to data include refreshed death information and additional cases that have been underwritten. Collaborating with a team of MDs, enhancements to the model address concerns identified upon individual case reviews. To date, new versions have focused on improving the accuracy of risk class recommendations for individual cases and specific medical impairments rather than aggregate performance. Prior to deploying a new version, we conduct a retroactive pilot to ensure no unexpected outcomes occur. The data science team generates new model outcomes for the past several months of cases, reports aggregate statistics, and the medical team analyzes individual model decisions before approval. Any change to the expected distribution of offers requires further approval from an actuarial team. Final deployment of a new version requires collaboration between data science and IT developers, who maintain the production system. The cadence for new model versions occurs on an as-needed basis, roughly biannually, rather than a scheduled frequency.

5.4 Regulation

The use of predictive modeling in life insurance underwriting raises legal, regulatory, and ethical questions related to transparency and fairness. In 2017, the New York Department of Financial Services requested that life insurers provide details of their use of algorithmic underwriting, including data sources, choice of model inputs, and the available mechanisms for disputing model-based risk decisions (Scism 2017). Further, the National Association of Insurance Commissioner's Model Rating Law requires underwriting inputs to be actuarially justified (i.e., demonstrate correlation with risk). Increasingly, insurers are being challenged to provide details around the inner workings of their underwriting models to provide both applicants and regulators a sense of which factors drive individual ratings.

A growing interest in consumer protection also manifests through concerns around fairness and the impact predictive models have on protected classes. The use of a wide variety of model inputs related to an applicant's personhood (e.g., age, gender, income) makes life underwriting models vulnerable to persisting societal biases that exist without the benefit of human manipulation to counteract its negative impacts on protected classes. In an effort to combat undesired biases, models and risk ratings are conditioned on certain protected classes, such as age and gender. In addition, purposeful omission of ethnicity and geography partially mitigates the risk related to fairness and disparate impact from use of algorithms in life insurance underwriting.

6 Business Value

The implementation of predictive modeling in life insurance underwriting has favorable implications for a firm's profitability and its customer experience. At MassMutual, the use of the mortality model and algorithmic underwriting has resulted in greatly improved operational efficiency—time to policy issuance has decreased by $>25\%$ for certain applicants. This improvement has had material impact on customer experience as indicated by a $>30\%$ increase of applicants opting to purchase their policies when the decision was made by the model compared to traditional underwriting within the best class. Further, the automation of underwriting decisions at the company has amounted to labor and time savings of millions of dollars in 2 years on a growing portfolio of policies that is valued in the tens of billions of dollars. Despite these operational financial gains, there is yet more profitability to be derived from the increased accuracy of the decisions when driven by the mortality model. That is, the retrospective simulation study detailed in Section 4.1 suggests a long-term benefit of reduced claims experience.

7 Conclusion and Future Directions

Pairing machine learning capabilities with historical data provides an unprecedented opportunity in the life insurance industry to transform the underwriting status quo. Leveraging 15 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.

Deploying a machine learning model and transforming a central business process has demonstrated the need to engage and collaborate with various partners beyond a data science team. Medical and underwriting have been crucial to improving the mortality model and its integration with the algorithmic underwriting system; actuarial and reinsurance stakeholders have vetted and approved a business-relevant evaluation framework; and legal partners have ensured that the process remains equitable in its treatment of applicants.

There are many avenues for future directions that span data, methods, and insurance innovation. The currently deployed mortality model does not consider all traditional underwriting data sources, such as prescription drugs or motor vehicle records, and there are non-traditional sources, such as financial data, public records, and wearable sensors, that may improve accuracy or enable alternative underwriting mechanisms. The general framework of producing high-resolution estimates of individual-level mortality risk can lead to actuarial and product innovation. Finally, trends in machine learning research on survival models may improve risk selection, and topics related to fairness and transparency of complex models are equally crucial to study.

8 Acknowledgments

The authors are grateful for contributions made by Paul Shearer, Martha Miller, John Karlen, Debora Sujono, and Sara Saperstein. We also give thanks to our many internal and external collaborators for their continued partnership.

References

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.

Atkinson, D. B., and Dallas, J. W. 2000. *Life insurance products and finance: charting a clear course*. Society of Actuaries.

Black, K., and Skipper, H. D. 2000. *Life and health insurance*. Prentice Hall.

Brackenridge, R. D. C.; Croxson, R.; and Mackenzie, R. 2006. *Brackenridge's medical selection of life risks*. Springer.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Case, A., and Deaton, A. 2015. Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century. *Proc. of the National Academy of Sciences* 112(49):15078–15083.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the Twenty-Second ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.

Chokshi, D. A.; El-Sayed, A. M.; and Stine, N. W. 2015. J-shaped curves and public health. *JAMA* 314(13):1339–1340.

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–c72.

Cox, D. R. 1972. Regression models and life-tables regression. *Journal of the Royal Statistical Society, Series B* 34:187–220.

Goldwasser, P., and Feldman, J. 1997. Association of serum albumin and mortality risk. *Journal of Clinical Epidemiology* 50(6):693–703.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests. *The Annals of Applied Statistics* 2(3):841–860.

Jordan, C. W. 1967. *Society of Actuaries' textbook on life contingencies*. Society of Actuaries.

Kalben, B. B. 2000. Why men die younger: Causes of mortality differences by sex. *N. Am. Actuarial Journal* 4(4):83–111.

Kaplan, E. L., and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):457–481.

Katzman, J.; Shaham, U.; Bates, J.; Cloninger, A.; Jiang, T.; and Kluger, Y. 2016. Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931*.

Kronmal, R. A.; Cain, K. C.; Ye, Z.; and Omenn, G. S. 1993. Total serum cholesterol levels and mortality risk as a function of age: A report based on the Framingham data. *Archives of Internal Medicine* 153(9):1065–1073.

Lipton, Z. C. 2016. The mythos of model interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, 96–100.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

Ranganath, R.; Perotte, A.; Elhadad, N.; and Blei, D. 2016. Deep survival analysis. *arXiv preprint arXiv:1608.02158*.

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 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

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):339–341.

Scism, L. 2017. New York regulator seeks details from life insurers using algorithms to issue policies. The Wall Street Journal.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*. 2503–2511.

Sugiyama, M., and Kawanabe, M. 2012. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT press.

Wilson, P. W.; D'Agostino, R. B.; Levy, D.; Belanger, A. M.; Silbershatz, H.; and Kannel, W. B. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18):1837–1847.

Wood, S. 2006. *Generalized Additive Models: An Introduction with R*. CRC Press.

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.