# RGA's Actuarial Validation of Milliman Irix® – Risk Score 3.0

## Mortality Risk-Based Predictive Models Using Prescription Drugs and Medical Billing Histories

**Hezhong (Mark) Ma** FSA MAAA
Vice President and Managing Actuary,
RGA Reinsurance Company

RGA

www.rgare.com

## Milliman Irix® Risk Score Versions, Data, and Model

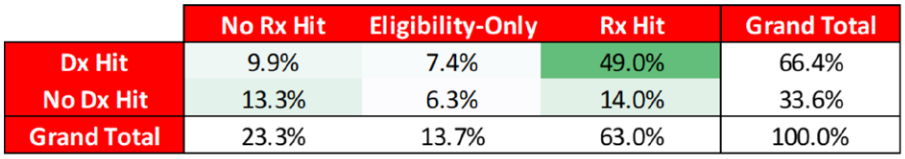
RGA's US Mortality Market provides unbiased assessment of risk segmentation tools. Our underwriting and pricing teams routinely engage clients and vendors in better understanding the industry trends and tools our clients use. Recently, RGA assessed Milliman Irix - Risk Score 3.0, which currently encompasses two mortality-risk-based predictive models. The first model, Risk Score 3.0-Rx ("Rx3.0"), utilizes prescription drug data ("Rx") only. The other, Risk Score 3.0-Rx & Dx ("RxDx3.0"), utilizes prescription drug ("Rx") and medical billing data ("Dx") when either or both are available. Milliman also provided RGA scores based on the Risk Score 2.2 version of the model ("Rx2.2"), to provide a point of comparison with the previous Rx-only model. No credit-related information is included in this analysis.

The dataset Milliman provided includes scores for 42.3 million individual insurance applicants spanning application dates between 2005 and the end of 2020. Deaths were also provided covering calendar years 2006 through the end of Q1 2021, allowing RGA to perform a mortality study with 236M exposure years and 1.7 million deaths. When compared to the validation data for the Risk Score 2.2 RGA received from Milliman, version 3.0 data has significantly more lives and deaths, especially for life insurance applicants. Similar increases are observed for older-age population and longer-duration exposure. The dataset identifies the line of business, such as life, final expense, health, etc. The life insurance line of business in the dataset represents a variety of underwriting methods, ranging from full underwriting to non-medical and simplified issue.

*RGA's US Mortality Market provides unbiased assessment of risk segmentation tools. Our underwriting and pricing teams routinely engage clients and vendors in better understanding the industry trends and tools our clients use.*

Not every individual has an Rx or Dx history scored. Exhibit 1 below illustrates the relationship between Rx3.0 and RxDx3.0. There are three types of Rx hits: those without Rx history (No Rx hit), those who exist in the enrollment data but do not have any Rx history (Eligibility only), and those with both enrollment and Rx history (Rx hit). There are only two types of Dx hits: those with Dx history and those without, as enrollment data was not provided for Dx. Both Rx and Dx have a relatively high independent hit rate. About 63% of exposures have an Rx history and therefore can be scored with both Rx3.0 and Rx2.2. For those exposures, RGA also received the Rx severity coding as red/yellow/green, for which red indicates more severe conditions that a medicine is meant to treat.  Eighty percent of exposures have either Rx or Dx history, and as such, can be scored by the RxDx3.0 model. In Exhibit 1 below, they encompass one of two components, the sum of the top row, 66.4%, and the exposure in the second row for those with Rx hits, 14.0%.

### Exhibit 1 – Summary of Exposure Received by Hit Types

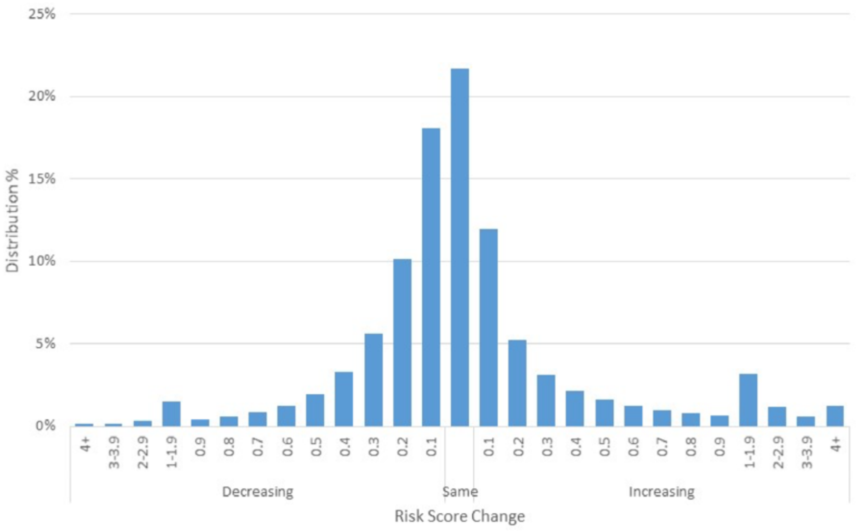


| | No Rx Hit | Eligibility-Only | Rx Hit | Grand Total |
|---|---|---|---|---|
| Dx Hit | 9.9% | 7.4% | 49.0% | 66.4% |
| No Dx Hit | 13.3% | 6.3% | 14.0% | 33.6% |
| Grand Total | 23.3% | 13.7% | 63.0% | 100.0% |

For the 14% exposure with Rx hit but no Dx hit, Rx3.0 score and RxDx3.0 scores are identical. Those with both Rx and Dx hits (e.g., the 49% of total exposure in Exhibit 1) could have very different Rx3.0 and RxDx3.0 scores. Exhibit 2 below measures the difference between Rx3.0 and RxDx3.0 scores for those exposure.

For about two thirds of those exposures, the difference between the two scores is less than 0.2. But some large differences do exist. About 6.2% of exposures see an increase of more than 1.0 from Rx3.0 to RxDx3.0. It is plausible that additional impairments are captured in the Dx history that are not in the Rx history. Additionally, 2.1% of exposures see a decrease of more than 1.0 from Rx3.0 to RxDx3.0.

Exhibit 2 – Comparing RxDx3.0 and Rx3.0 Scores for Those with Both Rx and Dx Hits



Both Rx3.0 and RxDx3.0 also ingest relevant demographic attributes. The Rx and Dx histories were grouped into clinically meaningful features. Machine-learning algorithms were used to predict mortality directly, as opposed to predicting underwriting decisions. The output of the model is relative mortality delivered as a numeric value ranging from 0.01 to 934.0

## Evaluation method

Based on the dataset from Milliman, RGA conducted an independent validation by developing a mortality study of the data through calendar year 2020. The expected basis throughout this article is the U.S. population mortality table by attained age, gender, and calendar year. Calendar year 2020 saw elevated mortality, but there is no attempt to adjust experience. Within the model training process, Milliman employed a train-test-validation split. There is no indication in the data for RGA to tell whether a person was used as training, testing, or validation. As such, the analysis in this paper is based on the entire dataset. It could be argued that the analysis based on the model validation data, instead of training/testing data during the model building process, is more representative of the results a carrier may see in production.

The result of the experience study is summarized as relative mortality by scores (e.g., lift curves). The steeper lift curve indicates the stronger power to segment good mortality risks versus bad risks. The lift curves of different models, or different business attributes of the same model, based on the numeric values, may not be directly comparable because of different distributions of exposure.

To control for this, RGA bucketed the experience study results into deciles. Each score bucket should always have 10% of the exposure among the subpopulation of interest. The same exposure might be decile 8 in Rx3.0 but decile 7 in RxDx3.0. This method helps separating the different segmentation power and the different distribution of exposures.

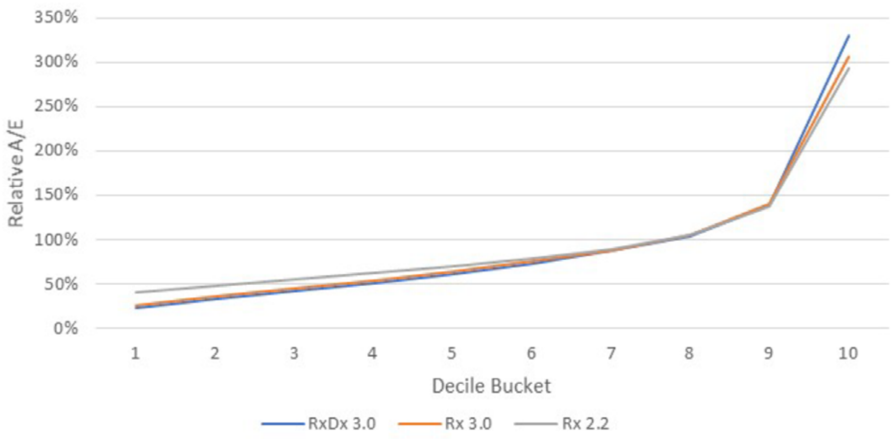
## Performance of Scores

Exhibit 3 presents relative mortality by decile buckets. Each data point represents 10% of the exposure for the model it represents. For example, the blue line represents the mortality lift curve of the RxDx3.0 model. As explained in Exhibit 1, 80% of the total exposure is scored by RxDx3.0. Therefore, each dot on the blue line represents 8% of the total exposure in the dataset RGA received. At the other side, the orange and gray lines are for those with an Rx hit regardless of Dx hit, about 63% of the total exposure. Each dot along those two lines represents 6.3% of the total exposure. Each line is adjusted to its own aggregate mortality level to derive relative mortality.

All three lines are monotonically increasing, indicating strong segmentation of mortality risks. RxDx3.0 has the steepest slope, suggesting the strongest segmentation power.

*All three lines are monotonically increasing, indicating strong segmentation of mortality risks. RxDx3.0 has the steepest slope, suggesting the strongest segmentation power.*

**Exhibit 3 – Relative Mortality by Decile Buckets**



The exhibit does not do justice to the models. The scales are dominated by the buckets with high relative mortality. The differences among the three lift curves are significant, even though they visually appear to be overlapping one another. Exhibit 4 presented the same data as Exhibit 3 but in a different format. Each model takes ratios to its correspondent Rx2.2 relative actual to expected ("A/E"). Naturally, the Rx2.2 turns into a flat line of 100%.

Exhibit 4 – Relative Mortality Using Rx2.2 as Baseline

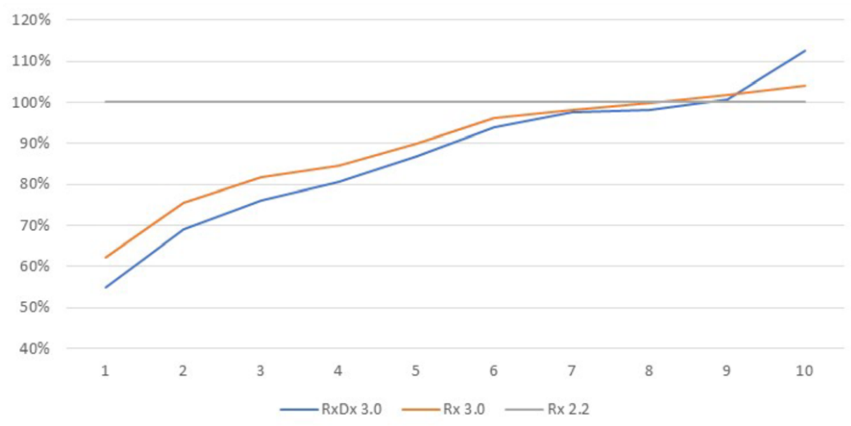


Exhibit 4 illustrates the improvement in mortality segmentation of Rx3.0 and RxDx3.0 from Rx2.2, across all decile buckets. From decile 1 to decile 6, Rx3.0 and RxDx3.0 can do a much better job in capturing exposure with lower mortality risks. At the higher end, Rx3.0 and RxDx3.0 do a much better job in capturing exposure with higher mortality risks.

Rx2.2 and Rx3.0 started from the same Rx history. The additional segmentation power comes from Rx3.0 being a more powerful model. In the first decile, the mortality of Rx3.0 is 62% of that of Rx2.2. Compared to Rx3.0, RxDx3.0 has more segmentation power from adding medical billing data into the model. In the first decile, the mortality of RxDx3.0 is 88% of that of Rx3.0.

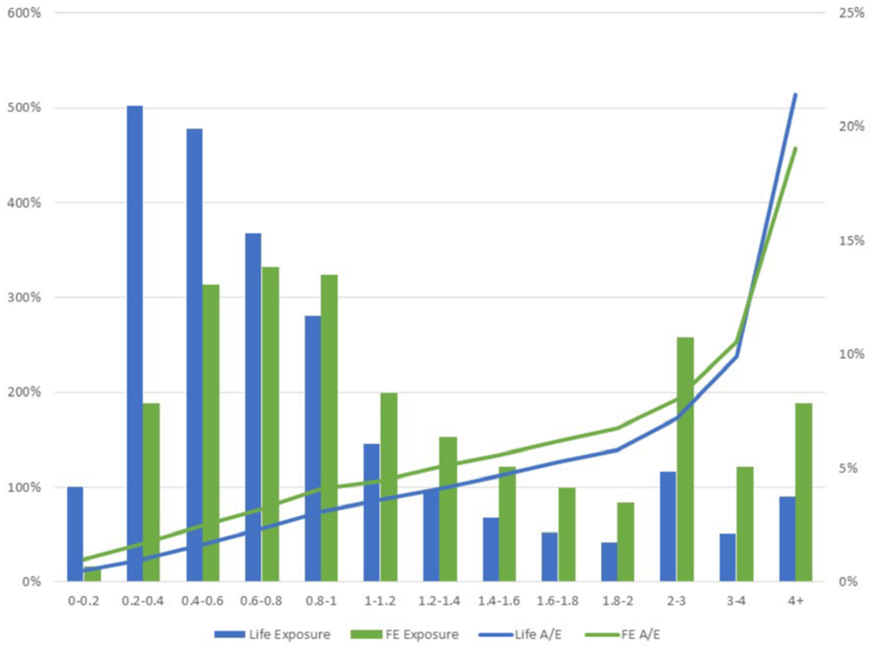
Another commonly used concept is lift, which is defined here as the multiple of the relative mortality of the top 10% exposure to that of the bottom 10% exposure. RxDx3.0 has a lift of 14.8 times. Rx3.0 has a lift of 12.1, while Rx2.2 has 7.2.

*Another commonly used concept is lift, which is defined here as the multiple of the relative mortality of the top 10% exposure to that of the bottom 10% exposure. RxDx3.0 has a lift of 14.8 times. Rx3.0 has a lift of 12.1, while Rx2.2 has 7.2.*

## Performance of RxDx3.0 by Lines of Business and Durations

As with all data products, performance and utility will vary by market segment, as well as attributes unique to each carrier that result in differences in the composition of their applicant pools. Context is always essential to understanding performance and utility. This section attempts to illustrate the different segmentation power of RxDx3.0 by different business attributes. Exhibit 5 looks at two different populations. The blue line represents a subset of the RxDx3.0 population from the life insurance line of business and without red severity drugs in their Rx history. The green line represents those from the final expense line of business. The Y-axis for those two lines represents the A/Es. The bars are the percentages of exposures within each score range and with colors corresponding to the business segments.  Note that the score ranges along the X-axis are grouped differently to produce a more even distribution of exposures. The first few are measured in increments of 0.2. Above score 2, the ranges are 2 to 3, 3 to 4, and 4+.

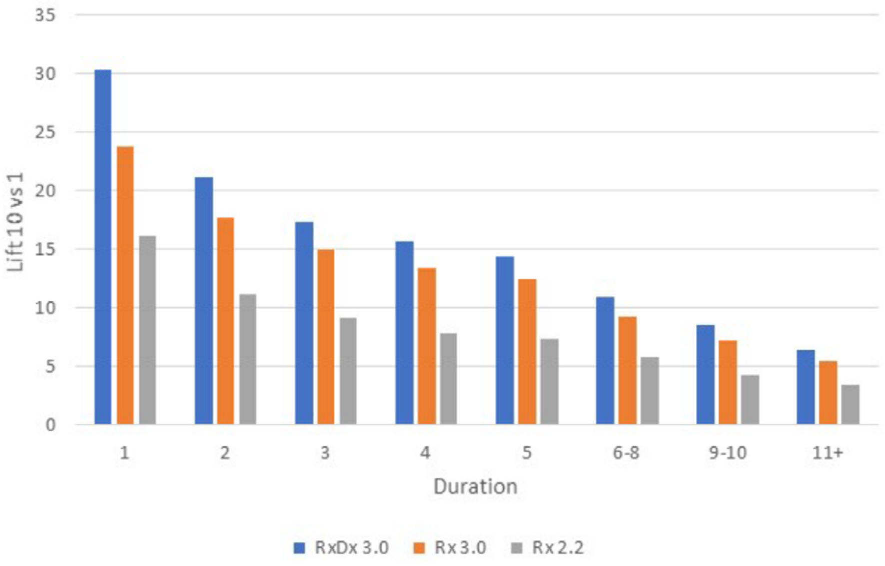
### Exhibit 5 – Relative Mortality by Lines of Business



Clearly, the two lines of business have different distribution by scores, as well as different levels and slope of mortality. Exhibit 5 highlights the performance differences based on the unique context associated with these populations. There is still the need for carriers to look at their own business mix to understand the impact of using scores for their products and use cases.

As with all underwriting evidence types, the favorable effects of underwriting wear off over time. Exhibit 6 illustrates the durational effects on lifts for Rx2.2, Rx3.0, and RxDx3.0. For this analysis, exposures were bucketed on the effective date, so a person would not switch buckets in later duration, even though the expectation is that given the high mortality associated with bucket 10, say, after a few years, there would be less than 10% of bucket 10 exposure left.

### Exhibit 6 – Mortality Lifts by Durations



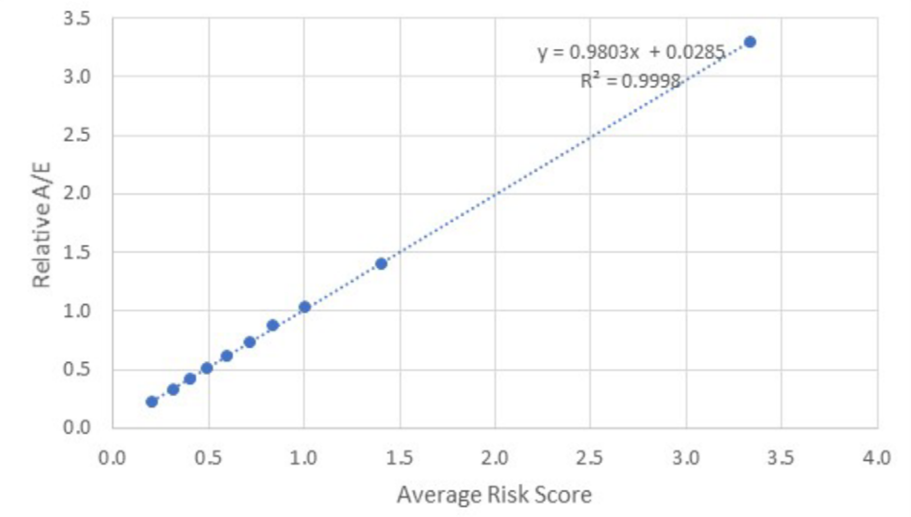
All three scores see wear-off of the mortality segmentation power as the duration increases. RxDx3.0 still demonstrates the greatest mortality segmentation power among all three after duration 11+ with a significant lift of 6.5 times. The lift for duration 11+ is 5.5 for Rx3.0 and 3.5 for Rx2.2.

## Mortality Meaning of the Scores

The output of the models is relative mortality. This is different from several other risk-based scores which rank risks but do not directly convey the level of relative mortality. As illustrated by Exhibit 5, where there is exposure with similar RxDx3.0 scores but different business attributes, RGA saw wide variations of relative mortality levels. But what about the proportional relationship? In other words, when scores double, does the mortality double?

Exhibit 7 presents the lift curve of RxDx3.0 in a different way. Each dot stands for one decile bucket. The X-axis is the geometric mean of the scores within that decile, while the Y-axis is the average A/E. A simple linear fitting is then used to find the slope of the line. It is 98%; very close to 100%.

### Exhibit 7 – Relative A/E vs. Average Scores



One interpretation of the slope of one is that if person A has a RxDx3.0 score twice of person B, person A’s mortality risk is about twice of that of person B. However, the slope varies slightly from population to population. For example, long-term-care sub-population has a lower slope than the other lines of business. If there is an interest in using this slope concept to derive mortality assumptions, it is important to understand the population of interest to ensure its unique characteristics are properly reflected.

## Summary and Limitations

Milliman Risk Score 3.0, both Rx3.0 and RxDx3.0, are effective in segmenting mortality. The newer generation of the scores outperform the previous generation product, Risk Score 2.2, in further segmenting both the low-risk end of the spectrum and the high-risk end. Including medical billing data into the RxDx3.0 leads to further segmentation when compared to Rx3.0, which is based on Rx history alone. Milliman Risk Score 3.0 can significantly segment mortality risks across the business attributes RGA examined, even though the level of segmentation varies by populations and durations.

This analysis is only as good as what the underlying data suggests. RGA did not have the detailed Rx nor Dx history to assess the reasonableness of scores based on the medical histories. In today's world, regulators and other stakeholders increasingly demand transparency, explainability, and fairness in using AI. Some carriers might desire more than a score to make an underwriting decision. An in-depth analysis could help.

Moreover, a carrier's application pool and insured population might not have the same underlying characteristics as the population in this study. As illustrated above, some measures vary by business attributes. While not considered in the study, the exclusivity is a crucial component of protective value, and the value of the scores will be impacted by other evidence used in underwriting. Therefore, RGA continues to see value in customized analysis to understand the impact of using scores for each company's products, market, and use cases. RGA is experienced in helping clients with deeper and more contextual analysis.

> *RGA continues to see value in customized analysis to understand the impact of using scores for each company's products, market, and use cases. RGA is experienced in helping clients with deeper and more contextual analysis.*