



IN BRIEF:

RGA's independent validation finds that Milliman's Irix® Risk Score 4.0 delivers stronger mortality segmentation and improved calibration accuracy compared with version 3.0. The analysis confirms that combining Prescription Data, Medical Data, and Credit Data enhances predictive power, while also highlighting structural changes in score distribution and performance across age, duration, and use cases.

RGA Validation: Milliman Irix® Risk Score 4.0

Mortality-risk-based predictive models using prescription drug, medical billing, and credit histories

By Laiping Wong-Stewart, Guizhou Hu, Hezhong (Mark) Ma

KEY TAKEAWAYS

- Milliman Risk Score version 4.0 more effectively distinguishes mortality risk across applicants and aligns predicted risk more closely with actual outcomes than version 3.0.
- Combining Prescription Data, Medical Data, and Credit Data consistently enhances mortality segmentation, with the most comprehensive scores delivering the strongest results.
- Changes in score distributions mean legacy cutoffs from version 3.0 may not translate directly, potentially affecting class size and risk selection if not adjusted.

Biometric risk represents the core uncertainty tied to human health and lifespan, which insurers must quantify and manage through underwriting, actuarial models, and portfolio design.

Introduction

Milliman's Irix® Risk Score (mortality) is a predictive model based on Prescription Data (Rx) that qualifies the relative mortality of life insurance applicants. The model's Rx-based input can be combined with Medical Data (Rx, Dx) or Credit Data (Rx, Cr) or both (Rx, Dx, Cr) to get a single, multivariate score. Milliman recently released Risk Score 4.0, which includes a Credit Data-only score (Cr). RGA conducted an independent analysis of the five risk scores' performance using data provided by Milliman.

Study Data

Milliman provided the dataset, which includes five types of risk scores for 57.6 million individual insurance applicants ages 18 to 85 whose application dates span 2005 to 2024. The data includes the following lines of business:

- Life (62% of the dataset)
- Health (11%)
- Final expense (7%)
- Medicare supplement (6%)
- Long-term care (3%)
- Disability (1%)
- Other (10%)

The cohort was followed through September 2024, yielding 383 million exposure years and 3.1 million deaths identified through the Social Security Master Index and third-party death records. Compared to RGA's previous series of white papers validating Milliman Risk Scores, we also received Credit Data-only scores, as well as the nicotine status flag, AnyNicotine. Milliman indicated that some or all of these datasets may have been used in score development, which should be considered when interpreting validation results.

Mortality was measured using the relative actual-to-expected (AE) ratio, with expected mortality based mostly on U.S. population tables, but certain versions of 2015 VBT tables were also evaluated. To account for underreporting deaths in the data, the analysis used relative mortality, or Relative AE, defined as raw AE divided by the total population AE.

Hit rate and mortality of no-hit

Due to the methodological shift in Prescription Data used in the study, Milliman stopped providing "Eligible only" as an Rx-hit type in the current dataset. Compared with the dataset used in [RGA's 2020 validation study](#) of Risk Score 3.0, both Rx (excluding eligibility only) and Dx hit rates increased in the current dataset.

Interestingly, the current dataset shows lower mortality in the Rx no-hit population than in the Rx-hit population. It may reflect changes in how Rx data was compiled between the two datasets. The new applicants starting in 2020 have a higher AE, most likely due to the COVID-19 pandemic. In that cohort, those with Rx hits have higher relative AE than those without. Some cases prior to 2020 were reclassified from no-hit to hit. Those populations likely have higher

mortality rates than those who were already a hit in the 2020 study. Carriers need to evaluate the expected hit rate of Rx in their production cases and the distribution of Rx histories and scores to make sure the mortality expectation is aligned to the population in this study.

Milliman ordered Credit Data on every adult for this study; therefore, “Credit Not Ordered” is no longer a hit type.

Table 1: Hit rate and no-hit mortality

	Hit rate	Relative mortality (with hit)	Relative mortality (no hit)
Rx	74.9%	101.2%	95.6%
Dx	69.6%	101.3%	96.0%
Cr	91.2%	97.1%	134.2%

Nicotine flags

For each case, Milliman provided a nicotine indicator, AnyNicotine, which is consistent with Irix Rules with ICGI = B. In the study, 8.6% of the population was flagged as ever-nicotine users. The AnyNicotine indicator was developed using a seven-year look-back period following receipt of the Medical and/or Prescription Data order from carriers. Milliman also noted that the study dataset included a lower prevalence of nicotine users than observed in production data, where more than 18% of Medical and/or Prescription Data queries include a nicotine indicator. Cases with a nicotine indicator have more than a two-fold increase in mortality, compared to those without the indicator. They also tend to have higher risk scores. Table 2 compares the average risk scores for cases with and without the nicotine indicators.

Table 2: Relative mortality and average risk score by nicotine usage

	Cases with AnyNicotine indicator	Cases without AnyNicotine indicator
Number of cases	4,923,591 (8.6%)	52,638,362 (91.4%)
Relative mortality	203.40%	92.60%
	Average Score	Average Score
4.0 score (Rx)	1.911	0.951
4.0 score (Cr)	1.366	0.936
4.0 score (Rx, Cr)	1.930	0.935
4.0 score (Rx, Dx)	2.282	0.918
4.0 score (Rx, Dx, Cr)	2.224	0.922

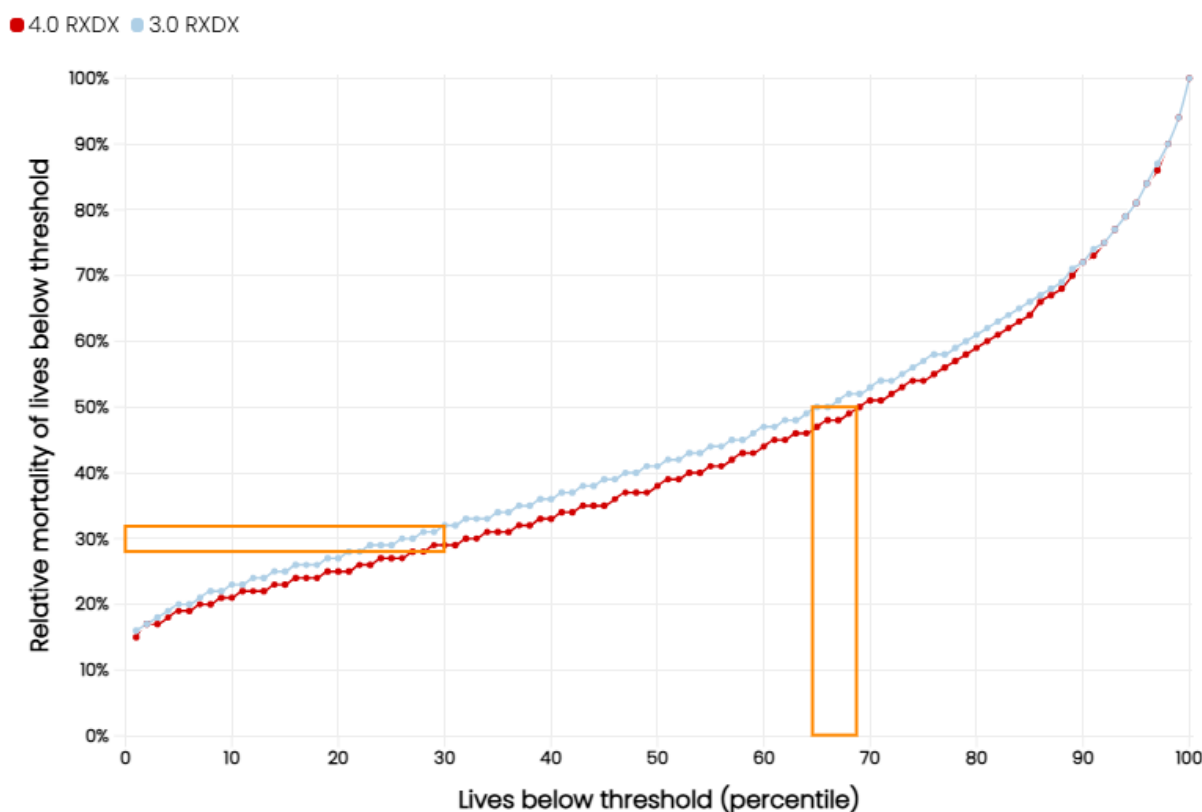
Assessing the mortality segmentation of risk scores

The primary purpose of these scores is risk selection and classification. In this context, a better-performing score is one that provides stronger mortality segmentation.

Figure 1 illustrates cumulative mortality by percentile for Risk Score 4.0 (RxDx) compared with Risk Score 3.0 (RxDx). Better-performing scores will have lower mortality at the same percentile. A horizontal comparison indicates that the 4.0 score qualifies a larger share of cases at the same mortality level. A vertical comparison indicates lower mortality for the 4.0 score at the same share of qualified cases.

For example, as shown in Figure 1, at a 30% STP target, cumulative mortality is 29% for Risk Score 4.0 compared with 32% for Risk Score 3.0. At a cutoff corresponding to 50% cumulative mortality, Risk Score 4.0 qualifies 67% of applications, compared with 64% for Risk Score 3.0. Not shown here, the analysis showed similar improvements from Risk Score 3.0 to their counterparts in version 4.0 for all five sets of risk scores (not just RxDx).

Figure 1: Relative mortality by risk score threshold



Score performance by sex and age group

Table 3 summarizes mortality segmentation with relative mortality at the 30th percentile across various scores by age group and sex, along with the improvement from version 3.0 to version 4.0. For illustrative purposes, the 30th percentile serves as an approximation to a cutoff

level that carriers may consider when identifying a preferred class. Scores that can be used to identify the best mortality risks in the top 30th percentile will have lower relative mortality in Table 3.

Table 3: Cumulative relative mortality by sex and age for the 30th percentile risk scores

		All	Female	Male	Age 18-35	Age 36-55	Age 56-65	Age 66-85
Score 4.0	Rx	34%	33%	35%	36%	30%	30%	44%
	Rx, Cr	28%	28%	28%	27%	25%	27%	42%
	Rx, Dx	29%	28%	30%	31%	26%	26%	39%
	Rx, Dx, Cr	26%	26%	26%	26%	23%	25%	38%
	Cr only	37%	38%	36%	35%	34%	39%	60%
Score 3.0	Rx	36%	35%	38%	39%	32%	32%	45%
	Rx, Cr	30%	30%	30%	28%	26%	28%	43%
	Rx, Dx	32%	31%	33%	35%	28%	27%	40%
	Rx, Dx, Cr	28%	28%	28%	28%	25%	26%	38%
Improvement of 4.0 over 3.0	Rx	2.4%	2.1%	2.5%	3.9%	2.1%	1.4%	0.7%
	Rx, Cr	2.1%	1.8%	2.5%	1.4%	1.4%	0.8%	0.7%
	Rx, Dx	2.7%	2.1%	3.4%	4.1%	2.0%	1.1%	0.6%
	Rx, Dx, Cr	1.8%	1.2%	2.4%	1.8%	1.3%	0.6%	0.4%

The key findings are as follows:

- Adding Dx to Rx, or adding Cr to Rx and Dx, improves mortality segmentation across all demographic groups.
- Score performance is broadly similar between males and females.
- All scores perform best in the 36-55 age group, followed by ages 56-65 and 18-35, and then worst in the ages 66+.
- The largest performance gain is observed in the 18-35 age group when Cr is added to RxDx, reducing relative mortality from 31% to 26%, a five-percentage-point improvement.

Notable improvements from version 3.0 to version 4.0 include:

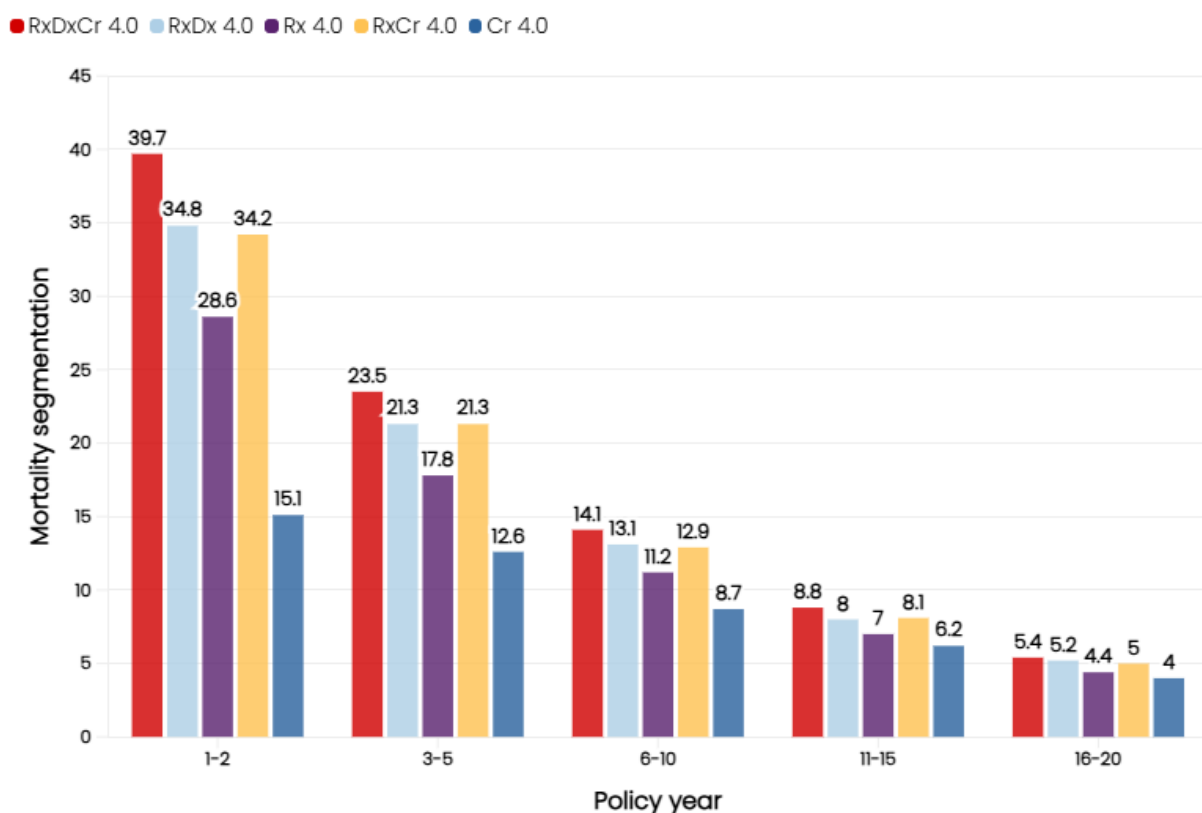
- All groups show improvement. This suggests that more experience and better models could increase segmentation power without introducing entirely new types of consumer data.
- Scores that already include Cr, such as RxCr and RxDxCr, show less improvement than Rx only and RxDx scores.
- Improvement is greater for males than for females.
- The largest gains occur at younger ages, with improvement declining as age increases.

Score performance by duration

In Figure 2, mortality segmentation is measured by the mortality ratio between the top and bottom deciles of the risk score. The mortality segmentation of all risk scores gradually declines with durations. The pattern of decline is similar for versions 3.0 and 4.0. These findings suggest that the interpretation of risk scores depends on duration.

Carriers should therefore use caution when applying the mortality segmentation observed in this study to their own populations, particularly when drawing pricing-related conclusions. It is worth noting that, even after 16 years, positive segmentation power remains evident.

Figure 2: Mortality segmentation (mortality ratio of 10th over 1st decile) by duration



Assessing the mortality prediction accuracy of risk scores

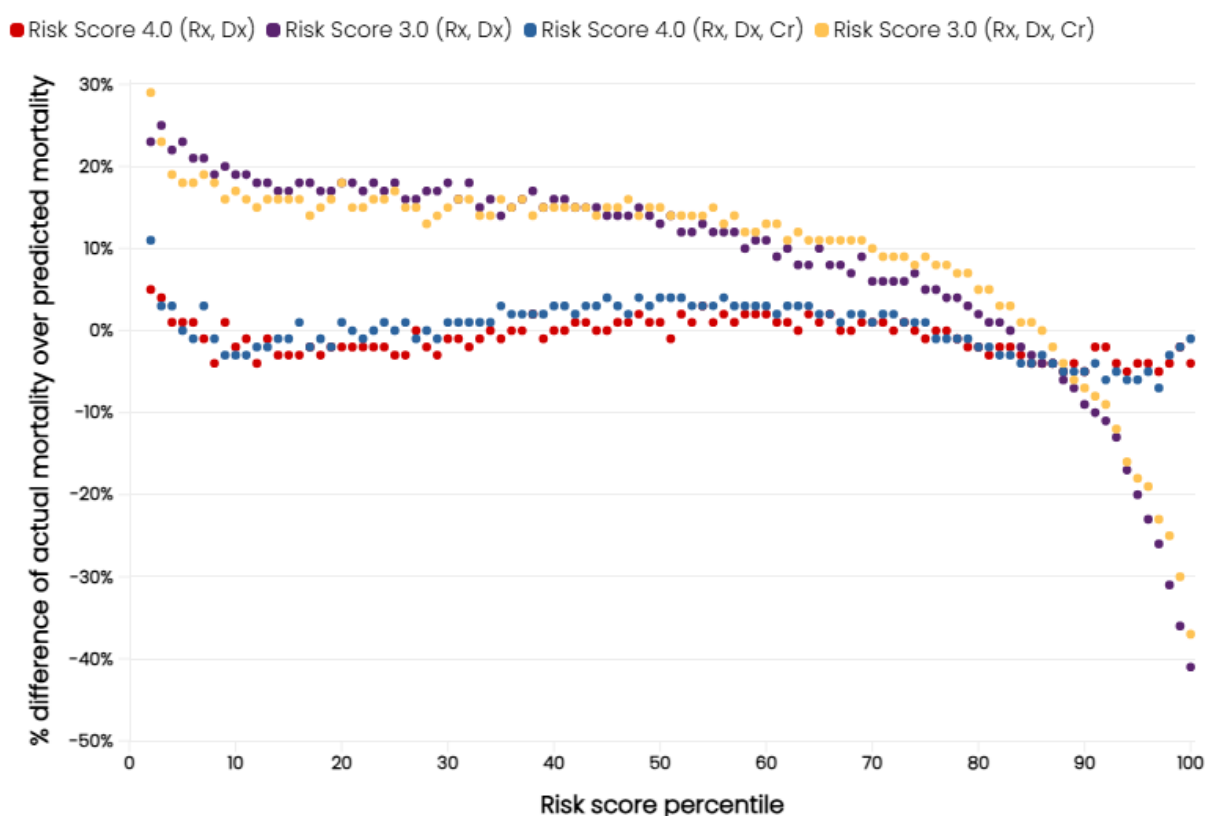
Risk score accuracy has two dimensions:

1. Discrimination accuracy, or mortality segmentation, which measures how well a score separates mortality risk
2. Calibration accuracy, which measures how closely predicted risk aligns with actual mortality

As discussed in the previous section, discrimination accuracy is particularly relevant for underwriting. Calibration accuracy, by contrast, is most relevant for making mortality inferences – for example, whether a 10% difference in risk score corresponds to an approximately 10% difference in mortality between two groups.

Figure 3 presents the calibration accuracy of two scores – RxDx and RxDxCr – across both versions, 3.0 and 4.0. It shows the percentage difference between actual mortality (relative AE) and predicted risk (the score) across the full score percentile range. When actual mortality is higher than the predicted mortality, the point lies above the 0% line; when actual mortality is lower than predicted mortality, the point lies below the 0% line.

Figure 3: Calibration accuracy of RxDx and RxDxCr for Risk Score 4.0 and 3.0



In Figure 3, both version 3.0 scores show points above the 0% line at lower-risk percentiles and below the 0% line at higher-risk percentiles, creating a downward trend. This pattern indicates that version 3.0 tends to understate mortality for lower-risk cases and overstate mortality for

higher-risk cases. By contrast, the version 4.0 scores are generally closer to the 0% line and appear relatively flat across percentiles, indicating improved calibration accuracy compared with version 3.0.

As noted above, the mortality segmentation of the risk scores declines with durations, which inevitably causes calibration accuracy to vary over time. Accordingly, the calibration results shown in Figure 3 should be interpreted in the context of the study dataset used in this analysis.

Score distribution shift between 3.0 and 4.0

Carriers currently using Score 3.0 and considering an upgrade to version 4.0 should recognize that applying the same score cutoff may lead not only to a different mortality impact, as discussed above, but also to a different class size because of the shift in the score distribution. Using a 0.3 cut point as an example, Table 4 below shows the percentage of cases that qualify at select cut points for the various scores. Note that the risk scores were rounded to the nearest tenth in comparison with the cut point.

Table 4: Percentage of cases with a score < 0.3

	Version 4.0	Version 3.0
Rx	13.4%	19.3%
Rx, Cr	20.1%	23.7%
Rx, Dx	18.7%	23.2%
Rx, Dx, Cr	23.0%	27.0%
Cr	11.8%	N/A

This pattern also highlights a broader difference in score behavior across versions. Adding more data to a score – for example, Dx to Rx or Cr to Rx Dx – generally widens the score distribution, allowing more cases to qualify at lower score thresholds. However, version 4.0 score distributions are more concentrated to the middle range than their counterparts in version 3.0. This reflects the refinement of version 4.0 to improve calibration accuracy.

Carriers should assess how a score version update may affect class size within their own portfolios. The percentages shown in Table 4 may not reflect a carrier’s actual business, as results can vary with differences in population mix.

Hypothetical business case

Many carriers use risk-based scores as rule-out criteria. If an applicant has a score over a certain cut point, the case can be declined or become ineligible for accelerated underwriting. This makes sense in that the mortality rate is generally low, and the overall mortality of a pool

can be determined by a small percentage of people with very poor mortality risks. However, that kind of use case does not apply the segmentation power of risk-based scores across the entire spectrum of the mortality risks.

Carriers can further use the segmentation power of these risk-based scores to separate preferred risks from a group consistent with a standard class in full medical underwriting. This analysis carved out a population with target mortality close to 100% of the 15 VBT AGG table to represent the standard class in full medical underwriting. The criteria are cases with ages 18–60, with a line of business labeled as life, no nicotine flags, and no Prescription Data flagged as “red.” Nicotine effects were assumed to be addressed separately.

For illustrative purpose, the analysis assumed the underreporting rates of death at 14%. With that assumption, the population with Rx Dx 4.0 below 0.7 has the mortality at approximately 100% of 15 VBT AGG. Further, the lowest 40% of the population in the study were considered preferred class. That translates to a cut point of 0.370. The mortality ratio of the preferred class relative to the standard class is 62%. That mortality ratio is within the range of reasonableness in terms of a typical life insurance underwriting program. Using risk-based scores could potentially help with preferred underwriting classification.

Certainly, this finding should not replace a retrospective study conducted by an individual carrier to fully assess how the score compares with conventional underwriting. RGA is experienced in assisting clients with deeper and more contextual analysis.

Summary and limitations

This analysis indicates that, across demographic segments, 4.0 scores generally differentiate between lower- and higher-risk lives more effectively than the prior version, although performance remains strongest in middle ages and weaker at older ages. Adding Dx to Rx, and Cr to Rx Dx, is associated with improved segmentation, while the AnyNicotine flag is associated with materially higher mortality and higher average risk scores.

The hypothetical business case further suggests that Risk Score 4.0 (Rx Dx) can identify a segment with mortality broadly consistent with a typical preferred class in a simulated population.

These findings should be interpreted in context. Score performance declines with duration, and calibration may vary depending on the duration mix of the population. Changes in data compilation also limit comparability with prior studies, including differences in hit types, hit rates, and the observed relationship between hit status and mortality.

In addition, score distributions differ meaningfully between versions 3.0 and 4.0, so legacy cut points may not translate directly to the new model.

The source data provided by Milliman may overlap with model development data and include limitations such as underreported deaths and assumptions used in the illustrative business case. Carriers should therefore evaluate performance and implementation impacts within their own portfolios before relying on these scores for underwriting or pricing decisions.