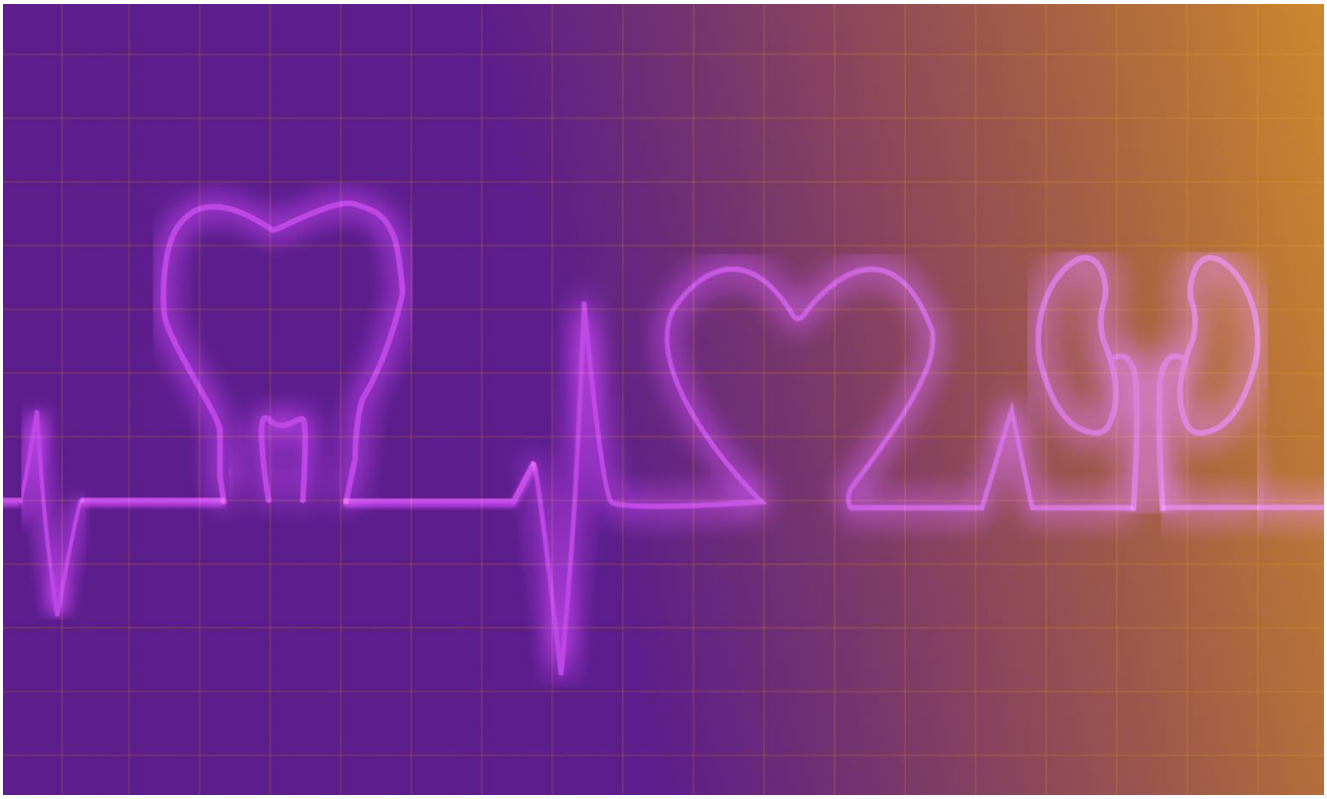


LECA Risk Score: A Predictive Mortality Model Using Individual's Dental Procedures History

Sikka Software, Sikka AI Lab (SAIL)
August 2022

Sikka.ai is the leading API Platform in the retail healthcare industry that includes dentistry, veterinary, audiology, optometry, chiropractic, orthodontics, oral surgery, and other medical practices. Sikka.ai API Platform seamlessly connects to 96% of the retail healthcare market throughout the US and is processing billions of transactions a day. Sikka.ai API Platform provides the underlying database of customer activities to help practices optimize their business, profitability, patient satisfaction, and patient medical history analysis, by gathering and analyzing their data into Sikka.ai's HIPAA compliant cloud in real-time. Since 2019, Sikka has been working with life insurance and reinsurance companies to perform retrospective analysis on millions of applicants with authorization and has refined the Life Expectancy at Current Age

(LECA) score, a predictive risk score based on a correlational analysis that measures a patients' mortality risk based on their dental procedure history.

Clinical studies[1] show a direct relationship between oral care and mortality rate, e.g. the number of teeth, dentures, implants, gum surgeries, extractions, periodontal diseases, and general oral care. One example is the relationship between Heart Disease, one of the leading causes of death in the US with about 630,000 deaths per year[2], and periodontal disease. Studies showed that the rate of gum diseases increases the risk of cardiovascular disease because gum diseases spread harmful bacteria and infections directly to the blood. When the harmful bacteria and infections reach a susceptible heart, they can cause inflammation called "Endocarditis". These studies were the initial trigger to perform a survival analysis study on Sikka's uniquely large dataset of dental records (currently at over 150M US patients) to assess the possible mortality risk correlation with an individual's dental procedures history. An actuarial analysis was performed on the type, frequency, and time-dependency of dental procedures and mortality rates. For prediction modeling, the Cox Proportional Hazard Regression was used to estimate the covariance coefficients, confidence intervals, and the statistical significance of the discretized attributes which are the frequency of distinct procedure codes across multiple calendar years.

Variable Importance Analysis

Survival analysis examines and models the time it takes for events to occur, termed survival time. The prototypical such event is death, from which the name "survival analysis" and much of its terminology derives. Survival analysis focuses on the distribution of survival times. The Cox proportional-hazards regression model is the most common tool for studying the dependency of survival time on predictor variables[3].

Over 18 million dental patients who visited a dental practice once or multiple times between 2003 to 2013 are ingested from the Sikka API Platform data under HIPAA compliant procedures. The US SSA Death Master Files are used to identify the number of deceased people during the same time frame (2003-2013). The US Life table on female and male mortality rates are from 2008 to 2013.[4] The result of the Cox PH model after trimming the non-significant variables (with p-value > 0.05) is shown in the ranked bar chart of Figure 1.

---

1 Relationship between oral health and mortality rate https://pubmed.ncbi.nlm.nih.gov/12472996/#:~:text=Results%3A%20A%20significant%20correlation%20was,indicator%20of%20all%2Dcause%20mortality.

2 National Center for Health Statistics. Health, United States, 2016: With Chartbook on Long-term Trends in Health. Hyattsville, MD. 2017.

3 "Cox Proportional-Hazards Regression for Survival Data in R." https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Cox-Regression.pdf. Accessed 6 Mar. 2018.

4 We are using the 2003 to 2013 life tables. Also have reviewed 2015 VBT tables. Our SMR plot is based on the US SSA Death Master files from 2013. https://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_03.pdf; https://www.cdc.gov/nchs/data/dvs/lewk3_2009.pdf; https://www.cdc.gov/nchs/data/nvsr/nvsr63/nvsr63_07.pdf; https://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_11.pdf; https://stacks.cdc.gov/view/cdc/42888; https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_03.pdf

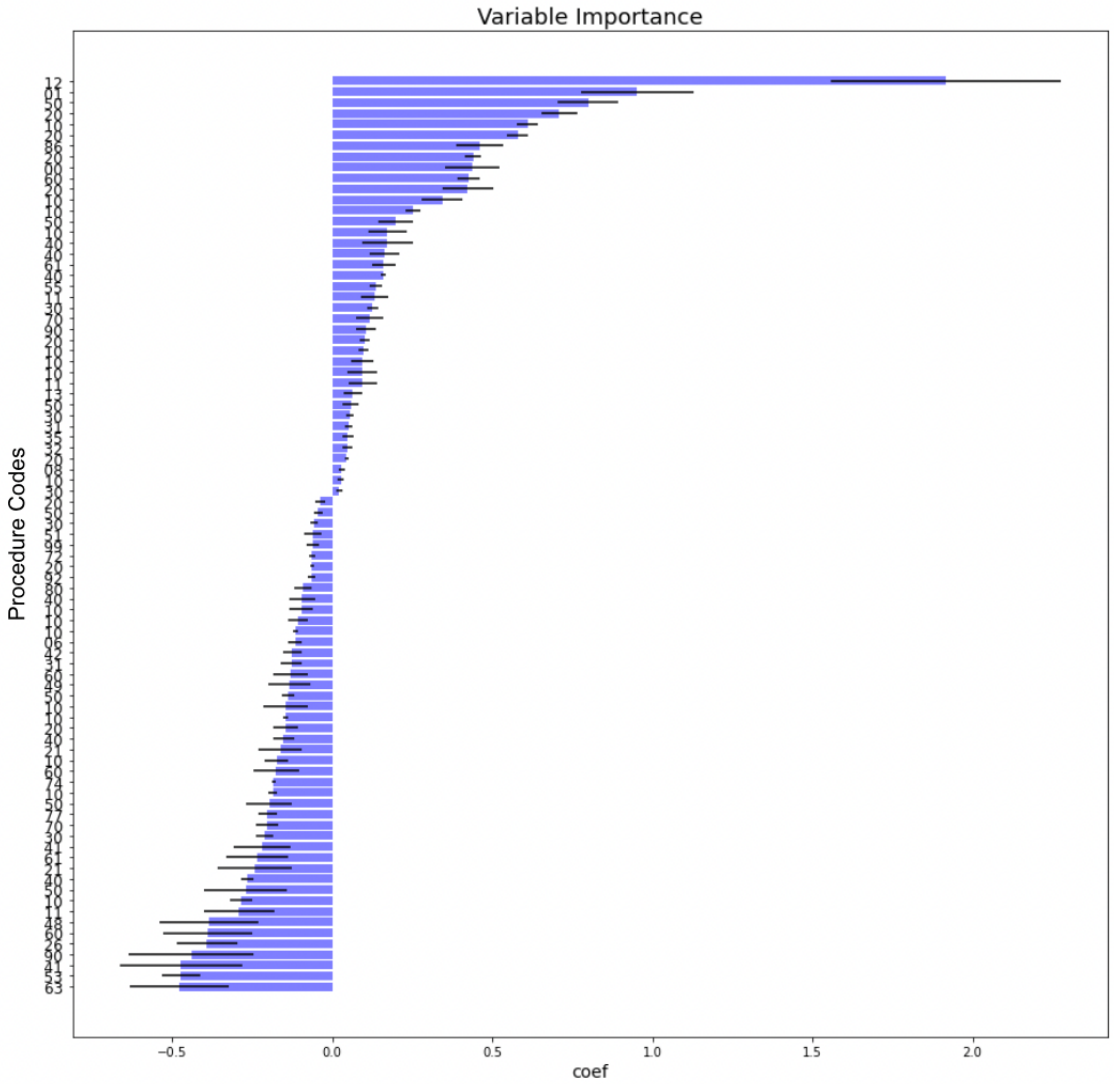


Figure 1. Variable Importance of Significant Dental Procedures

In Figure 1, the blue bar shows the coefficient of each significant variable with the black bar showing the confidence interval of each variable. The positive parameter indicates an increase in the relative mortality risk with respect to a one-unit increase in the value of the variable (e.g., heavy tooth and bone extractions, crowns, oral surgeries), holding other variables constant. With the same logic, the negative parameter reduces the mortality risk (e.g., periodic oral evaluation, preventive procedures). Overall, some dental procedures showed a statistically significant effect (p-value < 0.05) on the mortality rates to at least some degree.

Predictive Model

The survival analysis models were trained around 5 million patient records and the hazard ratios of each significant feature were used to deliver the LECA Risk Score (Life Expectancy at the Current Age Risk Score) for a given individual in the range of 0 to 2 with 0.1 intervals. The higher the LECA Risk Score, the

higher the mortality risk associated with that person. Sikka used a sample of around 4.8 million patients extracted from the Sikka API Platform Cloud as the test set. The trained model took the dental procedures as features and stratified based on 15-year age intervals (21-35, 36-50, 51-65, 66-80, >80).

Study Overview

In order to see how LECA Risk Score affects the survival rates over the following years, the test data set was split into 2003-2008 and 2008-2013. Individuals enter the study on January 1, 2008, and their LECA Risk Scores are calculated based on their dental procedures before 2008. The scores remain in the study until the date of death on records, or until the study ends (December 31, 2013). A lift relative mortality of the corresponding LECA Risk Score was observed between 2008 and 2013.

Key Terms:

"A" represents the actual deaths of the patients in the Sikka dataset who received dental care as identified in the US Social Security death master file from 2008 to 2013. To count an actual death, the social security number of the dental patient is found within the death master file.

"E" represents the expected deaths based on the changing age and gender mortality rates by years (2008-2013)[5]. Since on average, the mortality rate on Sikka platform is 3-5 times lower than the US population (see supplemental materials), the expected mortality rate each year is adjusted accordingly.

Example: If on average, the mortality rate at each age of the US population in 2008 is 3.2 times higher than that of Sikka platform of the same year, the expected mortality rate of 2008 will be divided by 3.2.

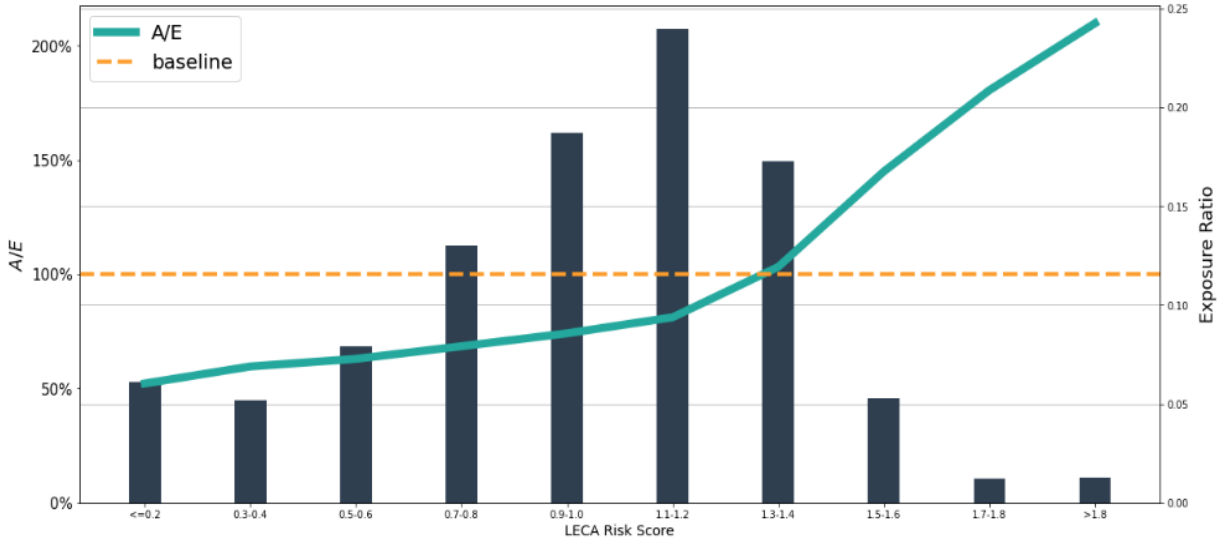
True expected mortality rate(age, sex, year)=Expected mortatlity rate(age, sex, year) / avg mortatlity multiples

"A/E" reflects the ratio of Actual / Expected deaths for each LECA Risk Score. It is also known as Standardized Mortality Ratio (SMR). As SMR increases, the actual mortality rates increase above the expected rates which shows the mortality risk stratification.

- A/E < 1 indicates there were fewer than expected deaths in the study population
- A/E = 1 indicates the number of observed deaths equals the number of expected deaths in the study population
- A/E > 1 indicates there were more than expected deaths in the study population (excess deaths)

5 2008, 2009, 2010, 2011, 2012, 2013 US life table. https://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_03.pdf; https://www.cdc.gov/nchs/data/dvs/lewk3_2009.pdf; https://www.cdc.gov/nchs/data/nvsr/nvsr63/nvsr63_07.pdf; https://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_11.pdf; https://stacks.cdc.gov/view/cdc/42888; https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_03.pdf

Mortality of LECA Risk Score

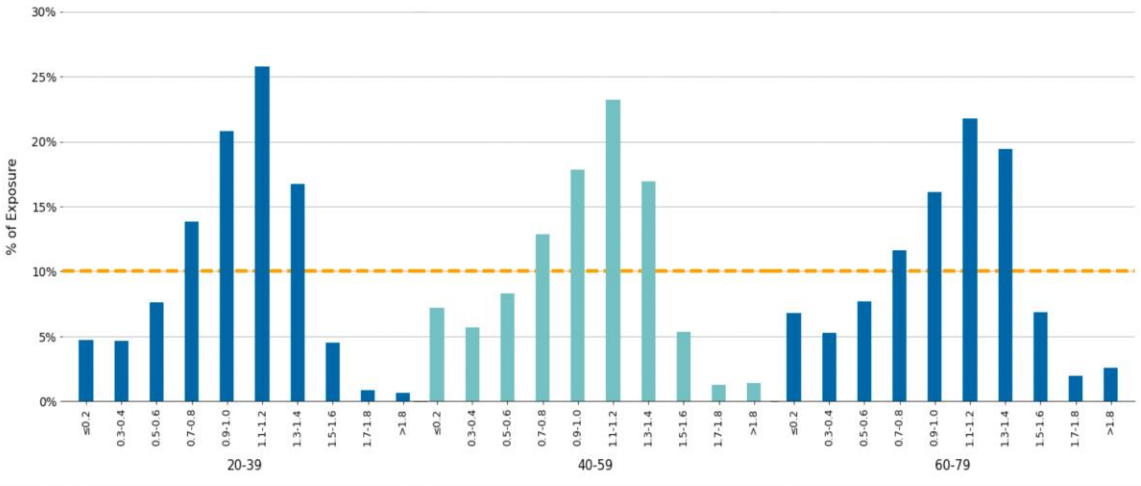


2008 Entry, 6 Year Follow Up, Entry Age 20-79

Figure 3. LECA Risk Score analysis and validation. Bars show the distribution of the LECA Risk Score bins with respect to the sample population in the study. Greenline shows the actual to expected mortality rate

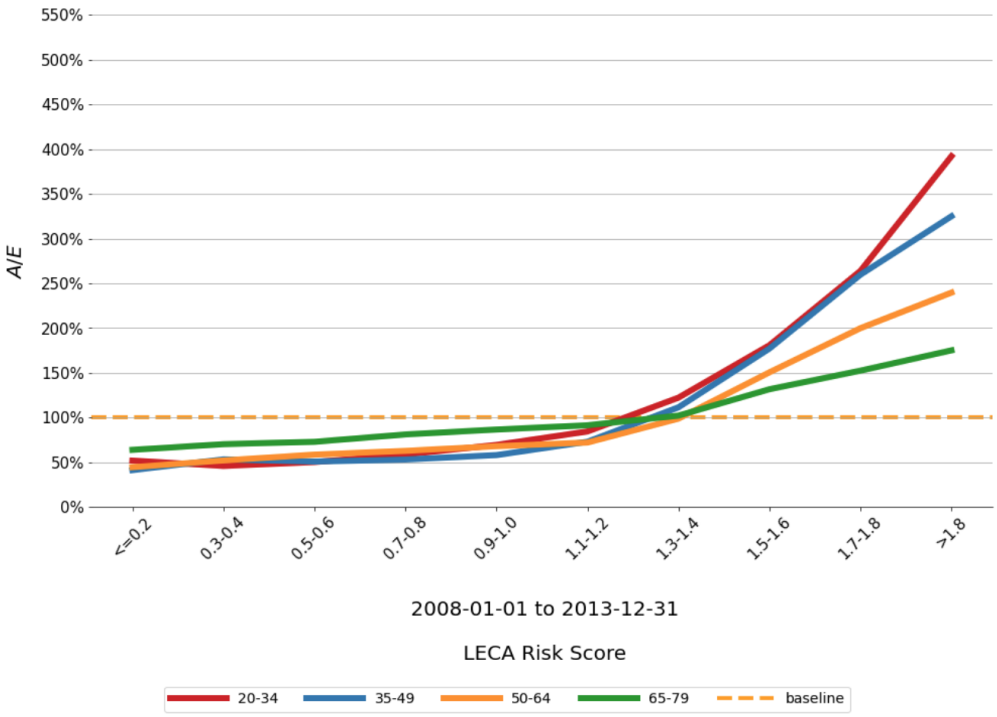
LECA Risk Score stratified relative mortality risk with a very strong agreement. The relative morality values visually have a semi-linear relationship with the LECA Risk Score that higher LECA scores contribute to a higher mortality rate. It is in fact fully consistent with the spirit of the LECA Risk Score coming from the log-likelihood ratios of the proportional hazard functions. The distribution of the LECA Risk Score shows a normal distribution with slight skewness. As shown in the graph, those having LECA Risk Scores higher than 1.8 were found to exhibit more than 4x the A/E ratio of those having LECA Risk Scores less than or equal to 0.2. Since age and gender are accounted for within the calculation of the expected deaths, the increase in A/E is due to the differences in dental procedures that are associated with the LECA Risk Score.

The distribution of LECA Risk Score depends on the individual's entry age. Figure 4 illustrates the score distribution of individuals entering the study on January 1, 2008. There is an observed high density of LECA Risk Score in the middle part for each age group.The proportion of low LECA Risk scores increases as age increases and flattens out after age 59.

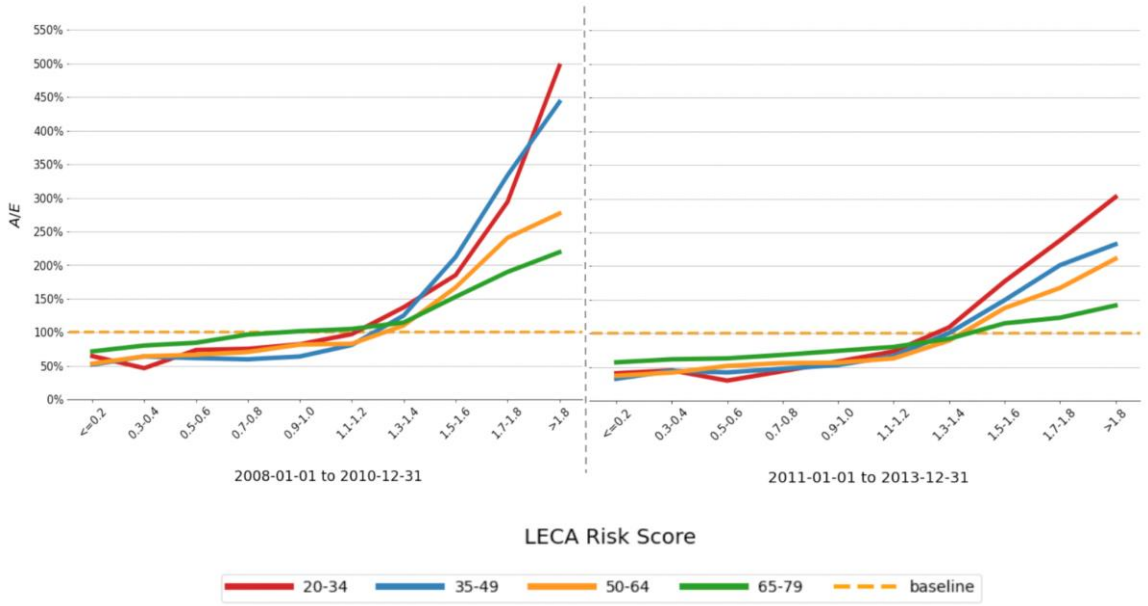


2008 Entry Year, LECA Risk Score by Entry Age
Figure 4. Distribution of LECA Risk Score by Entry Age

Figure 5 shows the A/E ratio by the different entry age groups, using 6-year follow up. For each age group, LECA Risk Score shows the power to differentiate the A/E across the different LECA Risk Score bins. In the higher LECA Risk scores, the actual deaths tend to exceed the expected deaths. Across all age groups studied, LECA risk scores provide significant predictive information regarding future mortality. In the age group 20-34 and 35-49, the highest LECA Risk Score bin has an A/E more than 7x that of the lowest LECA Risk Score bin.



2008 Entry Year, Entry Age 20-79
Figure 5. Actual to Expected Deaths by LECA Risk Score, Entry Age



2008 Entry Year, Entry Age 20-79

Figure 6. Actual to Expected Deaths by LECA Risk Score, Entry Age, and Study Period

Figure 6 analyzes LECA Risk Score's ability to segment risk for population at various study periods. Consistent mortality segmentations occur for different age groups but LECA Risk Score is more indicative of mortality outcomes in younger age groups. Notably, the predictive power of LECA Risk Score is still noticeable in each age group up to 6 years after scoring for the 20-79 demographic.

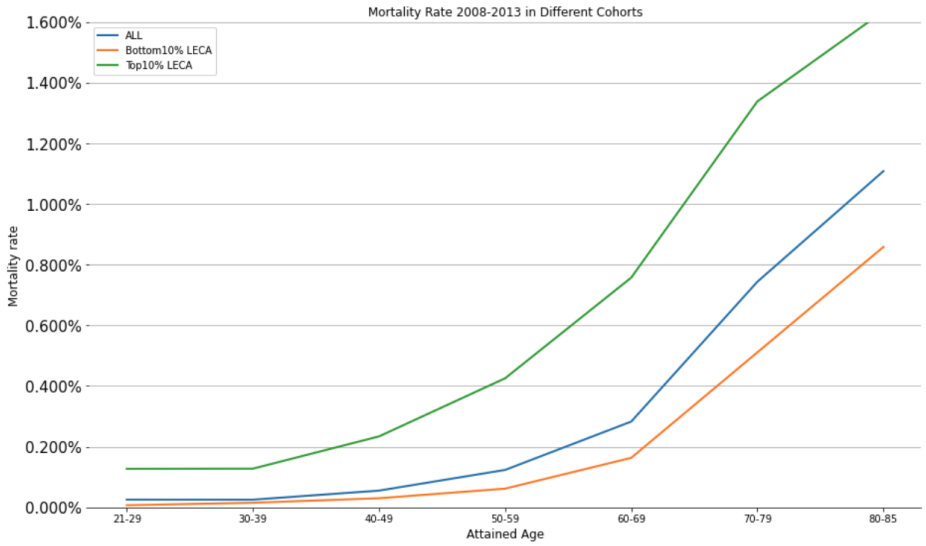


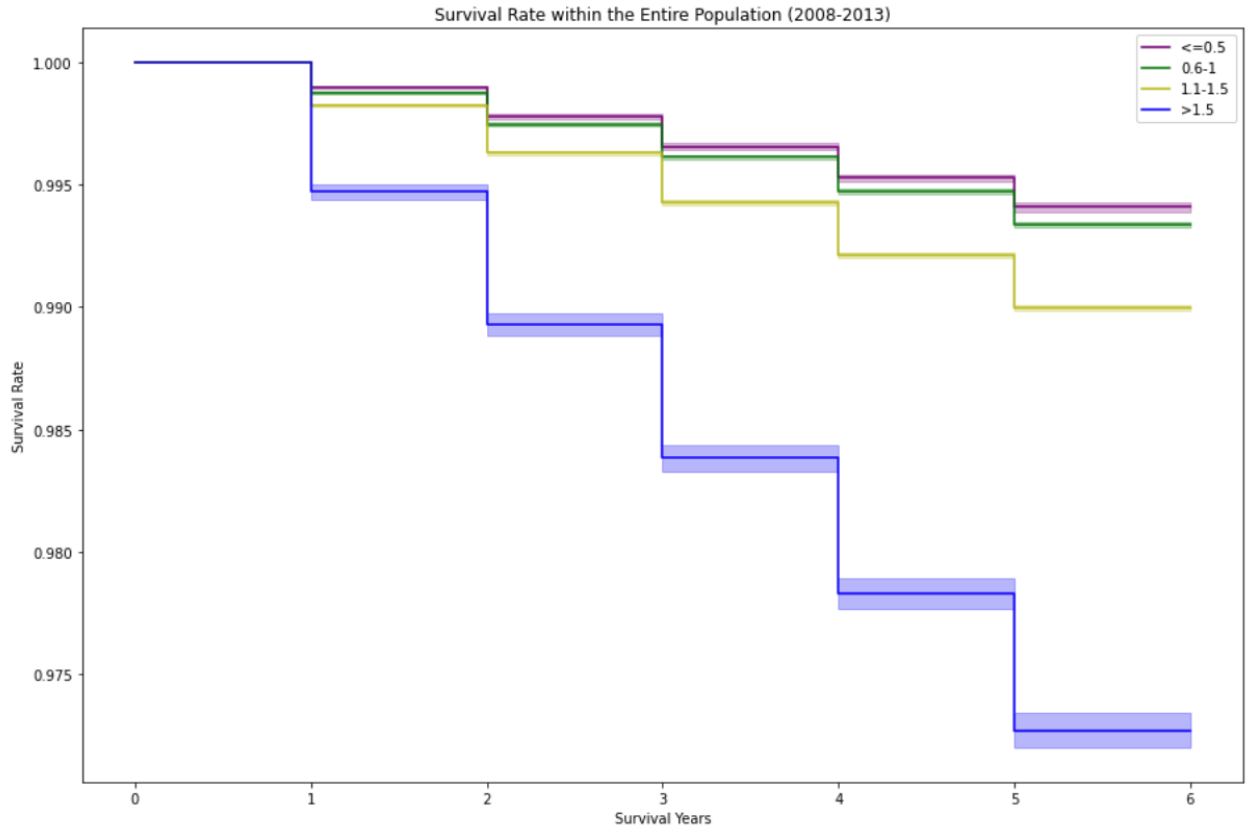
Figure 7. Mortality Rate by Top Bottom 10% and Overall Population, Attained Age

To see the magnitude of the 'lift' between the top 10% population and the bottom 10% population, we looked into the mortality rates by the attained age between 2008-2013 for three cohorts (Figure 7) - the overall population (baseline), the lowest 10% of LECA Risk Score for each attained age bins, and the highest 10% LECA Risk Score for each attained age bins. There is an observed higher mortality risk for

the highest 10 percentile of population compared to the overall population and the bottom 10 percentile of population.

Survival Rates of LECA Risk Score

The survival rate analysis was also performed to understand the predictive power of LECA Risk Score. Survival rate analysis was conducted on the entire test set between 2008 and 2013. Figure 8 shows that the survival rates drop faster with the high LECA Risk Scores in the entire population. Log-rank tests were also performed on the different LECA bins in terms of survival rate. The log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event (here the survival rate) at any time point[6]. If the p-value of the test is less than 0.05, we reject the null hypothesis and conclude that there is a significant difference between the two populations in the survival rate. It turns out each LECA Risk Score bin has a significantly different survival rate than the others.




```
<=0.5 VS 0.6-1  have significant difference in survival rate  p-value: 1.9521632066507825e-10
<=0.5 VS 1.1-1.5  have significant difference in survival rate  p-value: 1.8114249521682954e-223
<=0.5 VS >1.5  have significant difference in survival rate  p-value: 0.0
0.6-1 VS 1.1-1.5  have significant difference in survival rate  p-value: 1.2982295868756651e-287
0.6-1 VS >1.5  have significant difference in survival rate  p-value: 0.0
1.1-1.5 VS >1.5  have significant difference in survival rate  p-value: 0.0
```

Figure 8. Survival rate within the entire test set population vs. LECA Bins between 2008-2013 and Log-rank Test

[6] Log-rank test: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403858/#:~:text=The%20logrank%20test%20is%20used,of%20events%20(here%20deaths).

Figure 8-1 shows the survival rate between 2008 and 2013 vs. 10 LECA Risk Score bins by controlling different gender & age groups. The overall trend in each plot shows that the higher LECA Risk Scores have a faster drop in the survival rate in each different cohort.

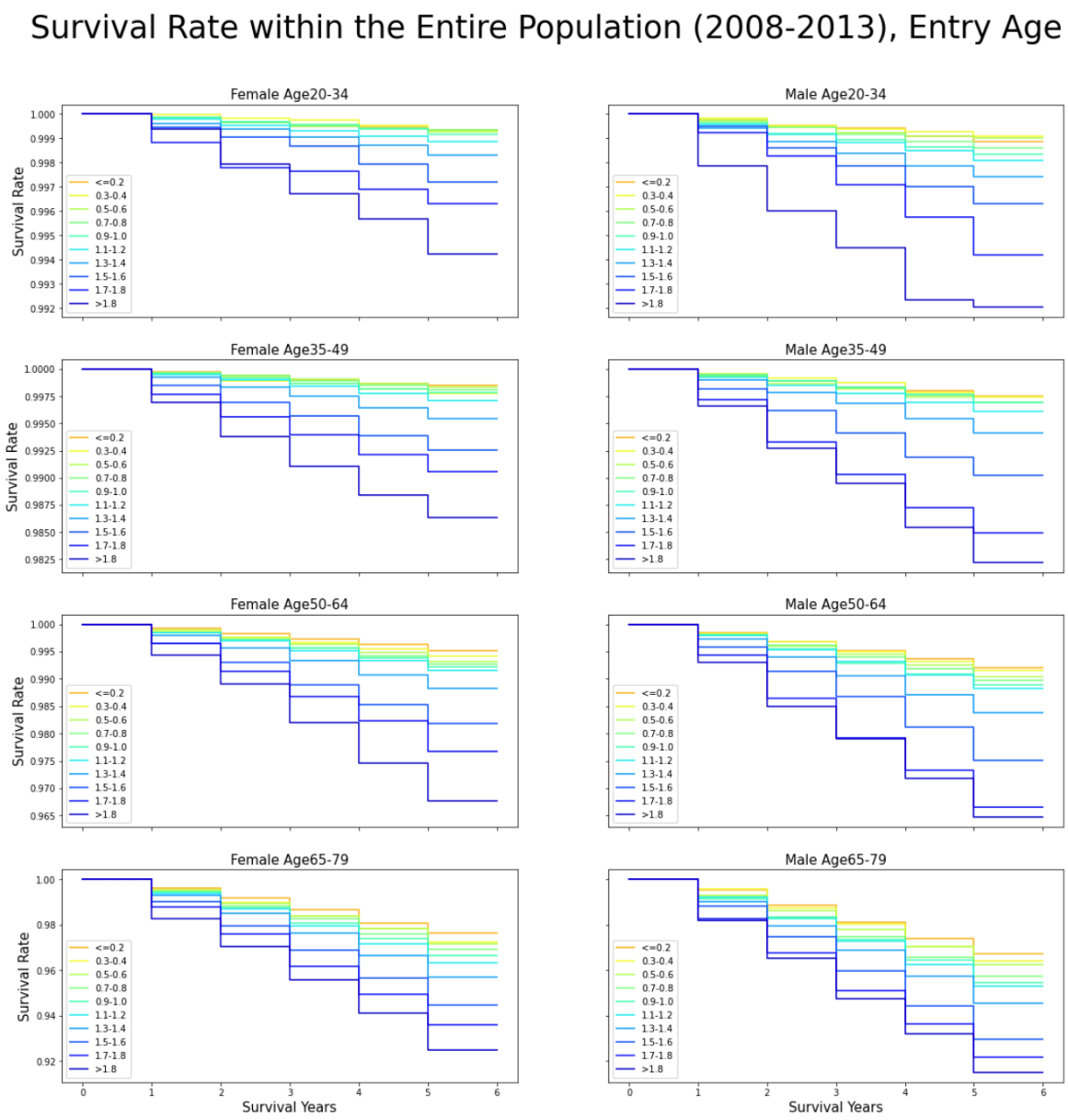


Figure 8-1. Survival rate within the different gender & age groups vs. 10 LECA Risk Score bins between 2008-2013

LECA's potential to improve underwriting

We set out to answer the question - "How would insurance underwriters incorporate another mortality-based score since they already have one based on age, gender, etc.?"

Underwriters would need to calculate the new information that LECA Risk Score could provide to their already existing age, gender, etc. based underwriting. For example, they may study the LECA results relative to their underwriting and determine that a 50-year-old male with a higher LECA Risk Score has a 30% higher mortality risk than a 50-year-old male with a lower LECA Risk Score. They could then create a table to adjust their underwriting based on the LECA Risk Score that factors in the higher morbidity into their rates.

Age	Gender	LECA<0.5	LECA>1.5
30	Male	0.98	1.20
50	Male	0.97	1.26
70	Male	0.92	1.40

A second question that we set out to answer was as follows - "The LECA Risk Score is solely based on dental procedures and not on existing underlying health conditions like cancer, diabetes, etc. How can it provide value for impaired applicants?"

We combined the disease data, age, and gender at the time of death with LECA Risk Score, to prove that the LECA Risk Score provides incremental information to predict mortality and doesn't duplicate other health data a life insurance underwriter would typically have from other sources such as the Rx database or EHRs. The below charts illustrate that a combination of the LECA Risk Score with other data sources[7] provides a lift in predicting mortality and improving the efficacy of a life insurer's underwriting process. Given that Sikka's LECA Risk Score is delivered in real-time via an API feed, it can help with automated or accelerated underwriting to better risk differentiate impaired applicants between standard and substandard tables, condition upcharges, and for Simplified Issue Term and other light underwriting products. Since we are offering an improvement to the existing underwriting models and trying to establish the efficacy of the LECA Risk Score in providing incremental value, we made several analyses on the actual mortality of disease data adding the LECA Risk Score to the same cohorts. This is because even simplified issue is going to have most of this information at the same time, and we are controlling for all of them (age/sex, tobacco, high blood pressure, diabetes, cardiovascular) in the comparisons.

[7] The disease data is ingested from Sikka Health Indicators following HIPAA compliant procedures. https://www.sikkasoft.com/Insurance

Figure 9 shows the mortality rate for high blood pressure patients and diabetes patients with various LECA Risk Scores. The plot below shows a positive correlation between the LECA Risk Score and the mortality rate for high blood pressure & diabetes patients. The yellow line shows the overall (average) mortality rate for high blood pressure patients and for diabetes patients in the test set. As the LECA Risk Score increases, the mortality rate in high blood pressure patients and diabetes patients also increase. When the LECA Risk Score decreases, the mortality rate in high blood pressure and diabetes patients reduces accordingly. The age at death is plotted for the same cohorts in three colors 20-39, 40-59, and 60-79. The gender at death is also added on top of the plots. The age at death of high blood pressure patients and diabetes patients and the gender distribution remain stable across different LECA bins in the plots.

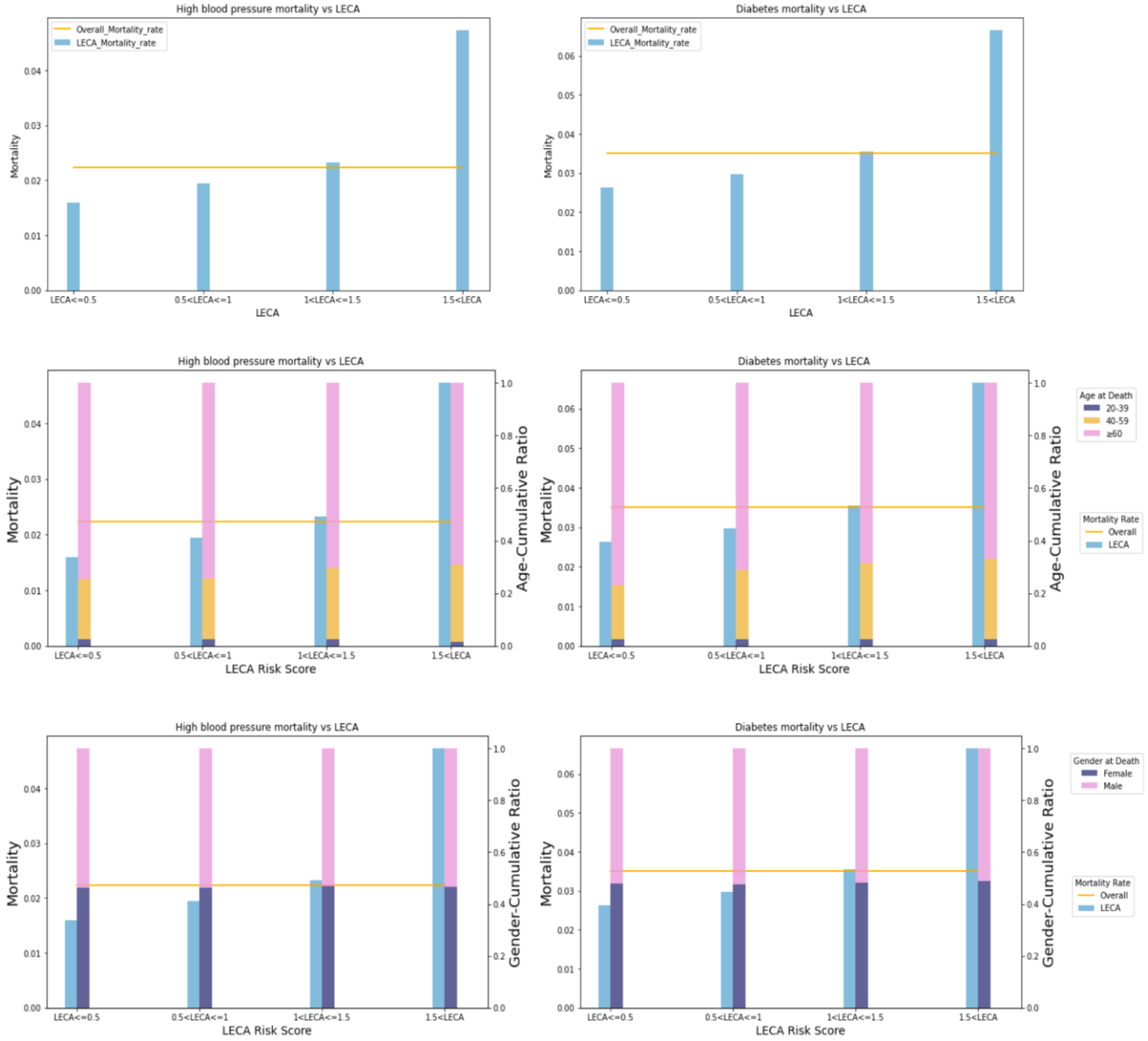


Figure 9. Mortality rate in the high blood pressure patients and the diabetes patients with respect to LECA Risk Score

Figure 10 shows the mortality rate in cancer and cardiovascular disease regarding the LECA Risk Score. The plot also indicates a positive correlation between the LECA Risk Score and the mortality for patients with either cardiovascular disease or a history of cancer. Age at death and gender at death are also added for the same cohorts.

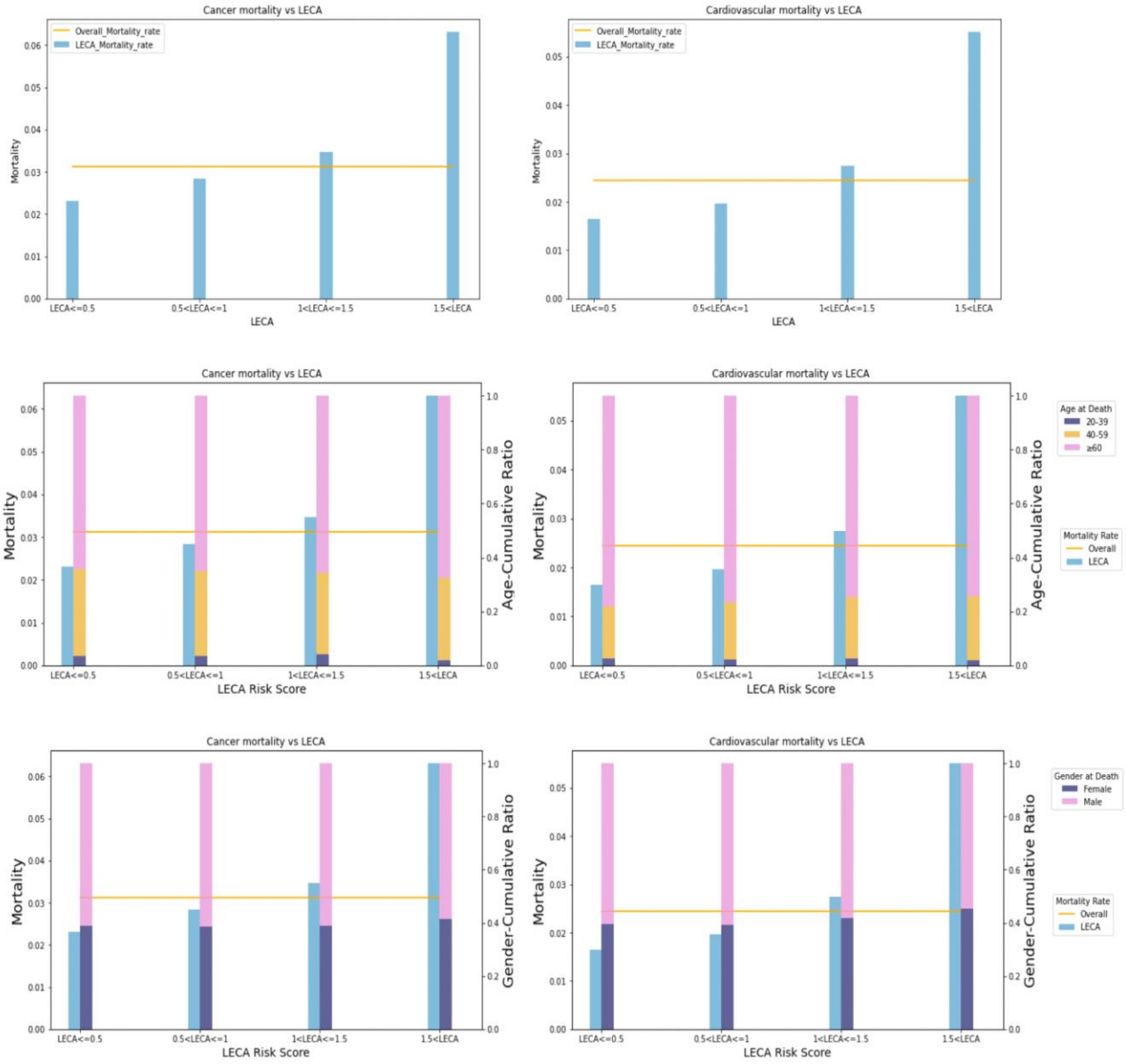


Figure 10. Mortality rate for cancer patient and patients with cardiovascular disease with respect to LECA Risk Score

Figure 11 shows Tobacco users' mortality rate vs. LECA Risk Score. We also can see the positive correlation between LECA Risk Score and Tobacco users' mortality. Age at death and gender at death distribution are also added in Figure 11.

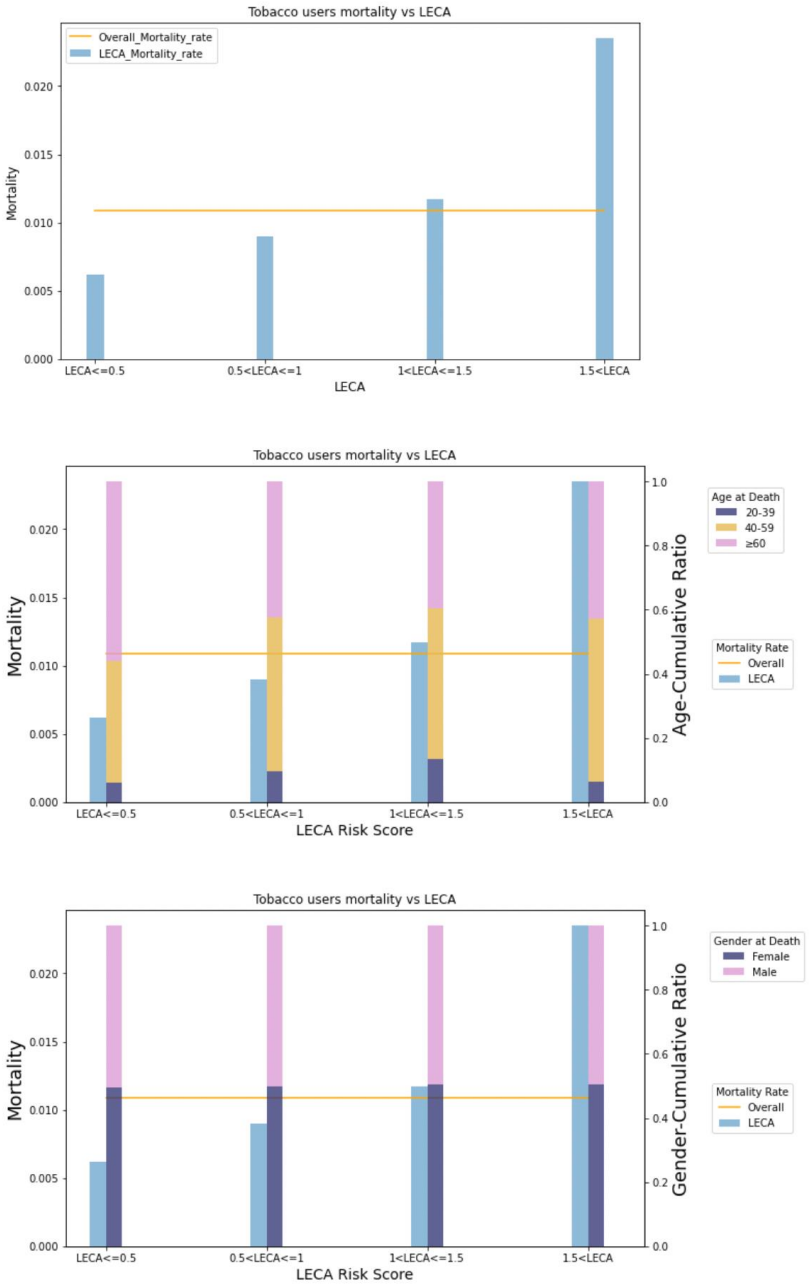


Figure 11. Mortality rate for tobacco users with respect to LECA Risk Score

The sample sizes of the health diseases used in the LECA Risk Score analysis are shown in table 1 below. The sample size for each indicator is based on around 4.8 million patients in the LECA Risk Score analysis after merging it with the US SSA Death Master records within the corresponding time period.

	Health Risk	Sample Size	Sample Coverage%
0	High Blood Pressure	265481	5.52
1	Diabetes	83103	1.73
2	Tobacco Users	88262	1.84
3	Cardiovascular	210895	4.39
4	Cancer	121443	2.53

Table 1. Sample size of high blood pressure, diabetes, tobacco users, cardiovascular, cancer

Figure 12 shows the mortality rates of patients with various diseases. The age at death of these combinations and the gender distribution also remain relatively stable in the plots. When LECA Risk Score increases, the mortality rates also increase in the patients who have these combination diseases.

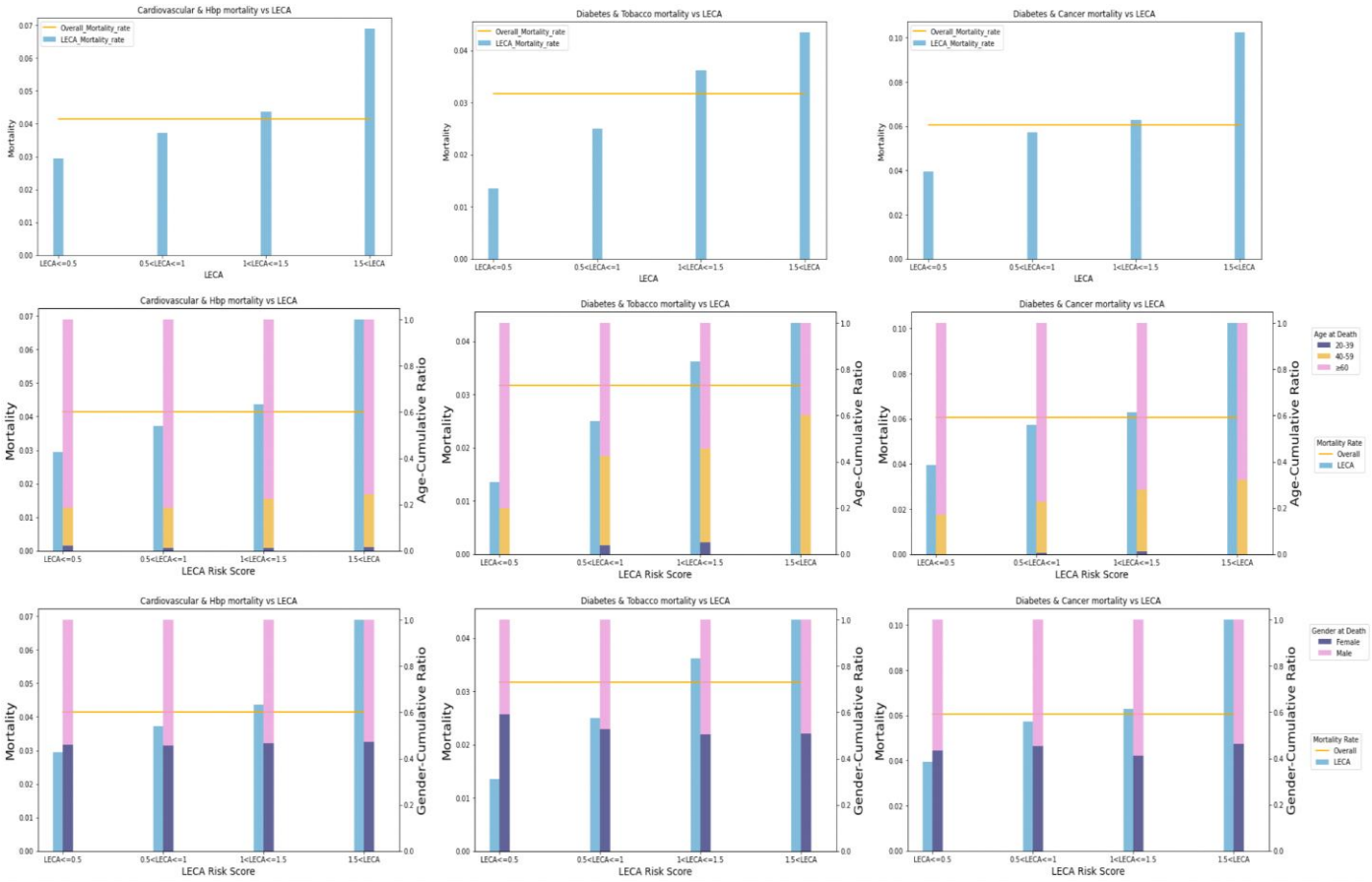


Figure 12. Mortality rate for patients who have various diseases with respect to LECA Risk Score

Suggested Uses for risk classification in life insurance underwriting including Substandard Risk Offers, upcharges for specific conditions, increasing the Automated Underwriting population, and table rating for Simplified Issue Term and other low underwriting life insurance products.

Sikka's research indicates a significant relationship between dental procedures & mortality risk, especially in certain impaired populations, which is reflected in the LECA Risk Score.

As the analysis above illustrates, the mortality rate of high blood pressure patients with a lower LECA Risk Score is lower than the average mortality rate for an average high blood pressure applicant. Similarly, the mortality rate for a tobacco user with a lower LECA Risk Score is lower than the average mortality rate for an average tobacco user. The same applied to diabetes patients. This indicates that an applicant who may be classified as substandard could be put into a standard policy with no deterioration in mortality loss by a carrier based on approving applicants with a LECA Risk Score of less than 0.5. In other words, a tobacco user who really takes care of their teeth and oral health may be at a relatively lower risk than a traditional smoker, and thus can be priced more competitively given the lower mortality rate.

Since the LECA Risk Score is available with the presence of appropriate authorization of an applicant within seconds, it makes it easier to use than a traditional EHR/EMR and suitable for a Simplified Issue Term product, other simplified issue products, and for any automated underwriting process with a high level of efficacy in terms of predicting relative mortality.

Summary

Sikka has found an interesting correlation between low LECA Risk Score and improved mortality despite chronic and habitual conditions. For example, a high blood pressure patient with a low LECA Risk Score or a tobacco user with a low LECA Risk Score has a lower mortality risk than the average high blood pressure or tobacco user. The same can be seen for diabetes, cardiovascular disease, and cancer. This use of oral care data can be very useful for a variety of underwriting situations including automated underwriting for a carrier, particularly in combination with other data sources as well as traditional age and gender tables. LECA Risk Score can be a powerful indicator in revealing mortality risk, especially in patients with some chronic conditions.

Sikka will continue to explore how insurance underwriters can incorporate the LECA Risk Score into their existing mortality risk models to improve the efficacy of underwriting and help make life insurance more affordable, especially for impaired applicants.

Supplemental materials

To see the generalization power of Sikka datasets, we compared Sikka data from 2003 to 2013 with U.S. population in two metrics: geographic and age distribution. The choropleth maps[8] below (figure 13) show the geographic population density of the U.S. residents and Sikka data from 2003 to 2013. Generally, in Sikka data from 2003 to 2013, states on the East and Southwest Coasts are slightly more populated than those on the U.S. map. Despite this, these two choropleth maps look similar to each other. The bar plots[9] below (figure 13) represent the age distributions of the U.S. and 2003 to 2013 Sikka data. The two bar charts are pretty similar. Thus we believe that age distribution in 2003 to 2013 Sikka data is akin to that with the U.S. population. In conclusion, Sikka data from 2003 to 2013 could be a geospatial representative of the U.S. population.

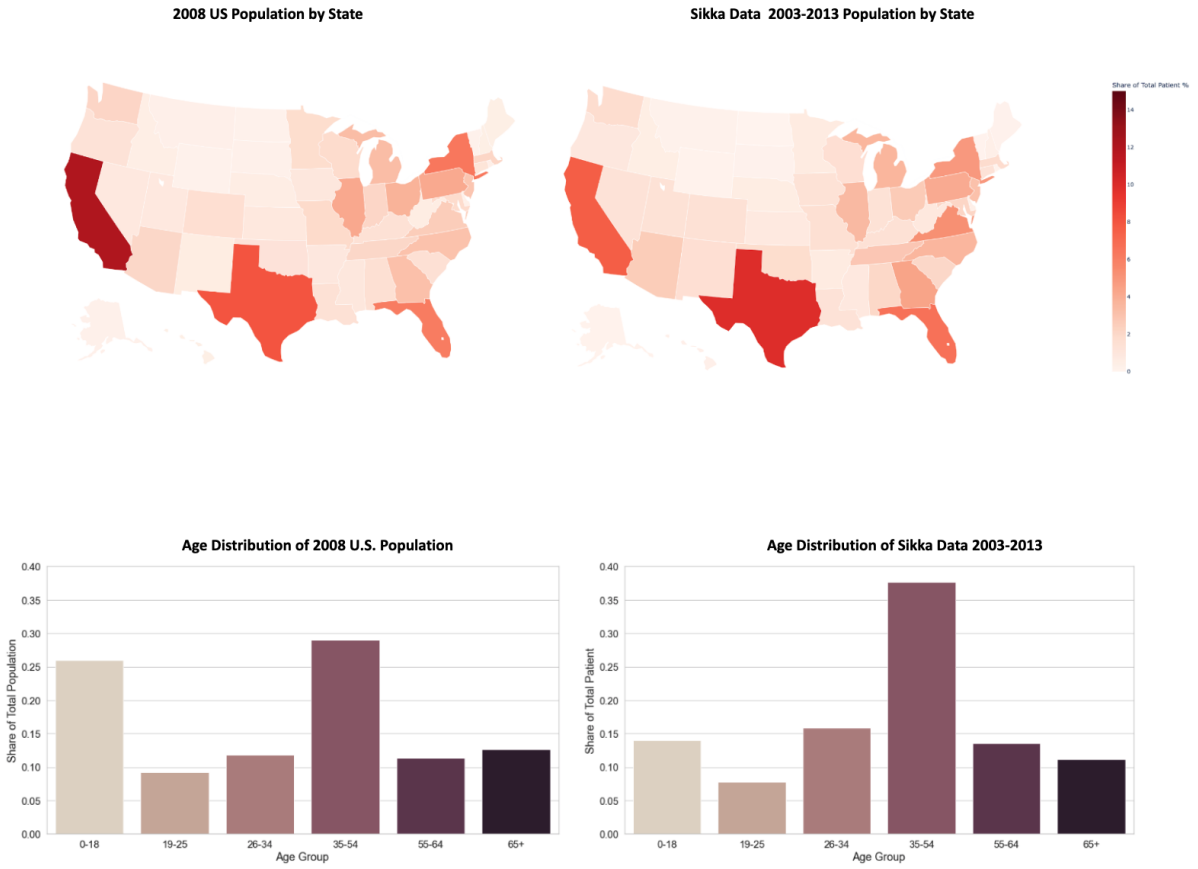


Figure 13. Population & Age Distribution in Sikka Dataset 2003-2013 vs. US 2008

8 Census Bureau, Population Division

9 U.S. Census Bureau, Population Division

Figure 14 compares the mortality rate of each age on Sikka platform with the US population. If we look at the mortality rate of the population aged 20-80, the entire data set from 2008 to 2013 has a lower mortality rate compared to the corresponding US population. It is observed that the average mortality rate each year of the US population is 3 - 5 times higher than that of the Sikka dataset. It is due to the fact that people who visit dental practices generally have a better financial and wellness condition (income level, access to dental insurance/provider, etc.).

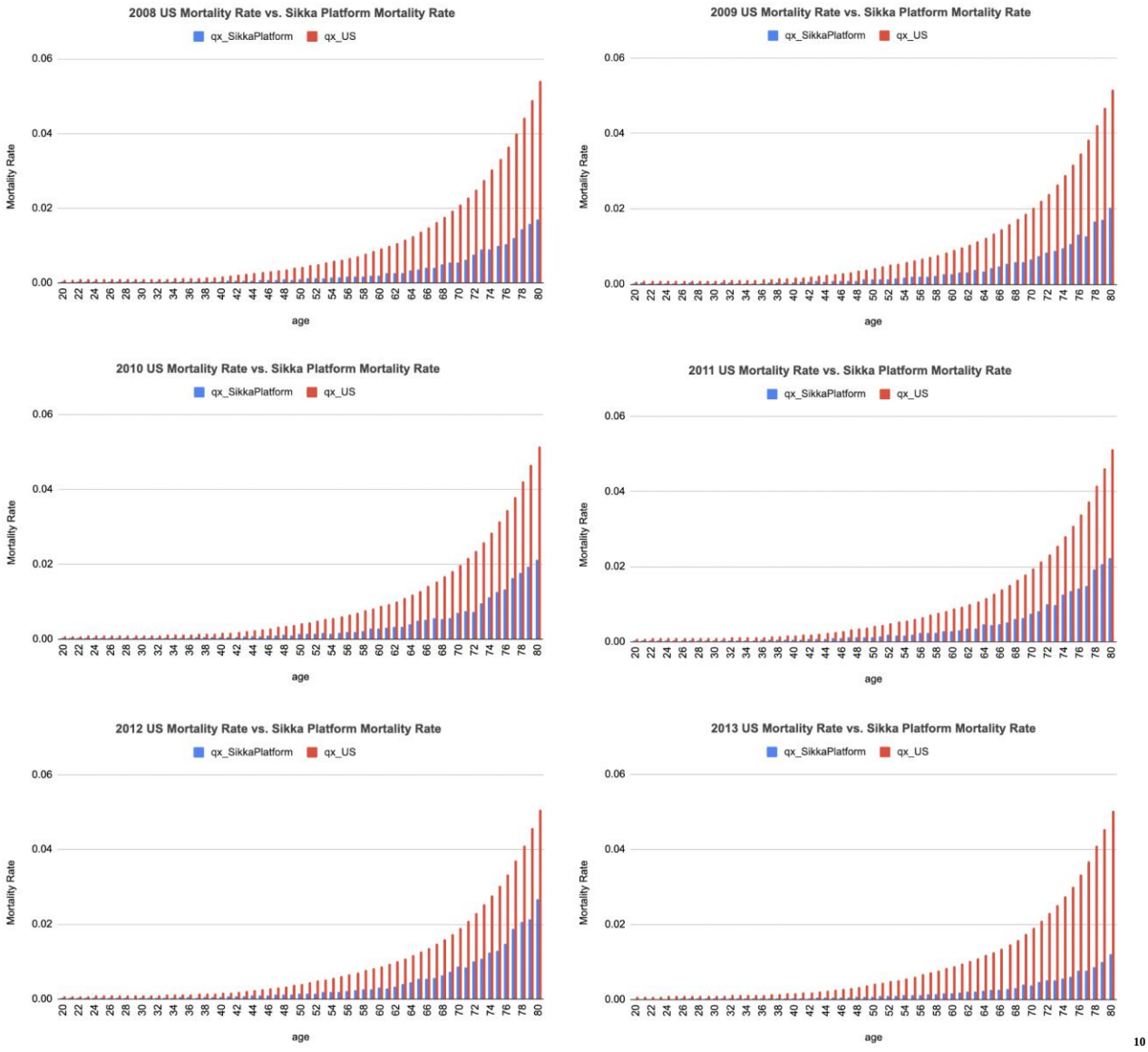


Figure 14. Mortality rate on Sikka platform between 2003-2013 vs. US population

[10] 2008, 2009, 2010, 2011, 2012, 2013 US life table: https://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_03.pdf; https://www.cdc.gov/nchs/data/dvs/lewk3_2009.pdf; https://www.cdc.gov/nchs/data/nvsr/nvsr63/nvsr63_07.pdf; https://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_11.pdf; https://stacks.cdc.gov/view/cdc/42888; https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_03.pdf

References

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression, 2002*.

Nicholls, Anna. “The Standardised Mortality Ratio and How to Calculate It.” Students 4 Best Evidence, 21 Sept. 2020, s4be.cochrane.org/blog/2020/08/26/the-standardised-mortality-ratio-and-how-to-calculate-it/.

“Social Security.” Actuarial Life Table, www.ssa.gov/oact/STATS/table4c6.html.