SOCIETY OF ACTUARIES

Mortality and Longevity

Incorporating Datavant's Death Index in Mortality Analysis

November 2020

Incorporating Datavant's Death Index in Mortality Analysis

AUTHOR

Mingke Du
Jun Hou
Benjamin Hsu
Yun Tang
Yubo Wang
Abraham Weishaus
Lina Xu

Columbia University

SPONSOR

Society of Actuaries

CONTENTS

Executive Summary **4**

Section 1: Datavant's Death Index (DDI) **4**

1.1 Data Features 4

1.2 Data Reliability 6

1.2.1 Record Selection 6

1.2.2 Death Counts by Year 6

1.2.3 Death Counts by Age 7

Section 2: Incorporating Datavant's Death Index in Mortality Analysis **8**

2.1 Mortality Improvement Model 8

2.2 Mortality Rates with DDI Death Counts 9

2.3 Mortality Improvement Rates 10

Section 3: Acknowledgments **12**

Appendix A: Death Counts from Datavant's Death Index **13**

Appendix B: Mortality Improvement Rate Updates **15**

References **19**

About the Society of Actuaries **20**

Incorporating Datavant's Death Index in Mortality Analysis

Executive Summary

One of the evolving datasets with the COVID-19 Research Database is the Datavant population mortality information. This report examines the efficacy of incorporating Datavant's Death Index in U.S. population mortality analysis. Mortality data is provided by Datavant and contains obituary data sourced from online newspapers, funeral homes, online memorials, direct submissions and more. The mortality database contains very current individual death information not concurrently widely available in other databases and is available to researchers on the COVID-19 Research Database platform. The data, technology, and services used in the generation of these research findings were generously supplied pro bono by the COVID-19 Research Database partners, who are acknowledged at https://covid19researchdatabase.org .

This report notes the features of the database, discuss their potential use and limitations, examines data reliability and reviews it potential use as an additional data set and input to mortality improvement models. While Datavant's Death Index captures deaths unevenly across age and under-captures population deaths, its up-to-date information on individual deaths, when appropriately aggregated and modeled, can be informative of recent mortality trends.

Section 1: Datavant's Death Index (DDI)

Datavant, Inc., a private U.S. company specializing in organizing health data, has recently made its mortality data available to researchers on the COVID-19 Research Database platform. Datavant's Death Index is a mortality database with information on deceased individuals in the United States and Canada. It contains information on individual deaths and is updated on a weekly basis. The database sources information from the Social Security Administration's Death Master File and also from public and private obituaries since 2010. Its up-to-date information, not concurrently available in other databases, can be aggregated and modeled for analysis.

In this section the reliability of the Datavant's Death Index (DDI) is examined. First, the available data in this database is highlighted and leads to a discussion of any potential limitations. Death counts are then aggregated from the DDI and compared to those in the Human Mortality Database (HMD).

1.1 DATA FEATURES

Each record on a deceased individual contains general information and encrypted personal information. Items containing general information, such as dates of birth and death, are referred to as "filters". Combinations of personal information (such as name, date of birth, and gender) are encrypted into "tokens" to protect the individual's privacy and serve as a means to match individuals across datasets. A sample record is shown in Figure 1.

Figure 1
SAMPLE DDI RECORD

OPERATION	DEATH_DATE_IMP	DATE_OF_DEATH	DATE_OF_BIRTH	GENDER_PROBABILITY	GENDER	TOKEN_1	TOKEN_2	TOKEN_4	TOKEN_7	TOKEN_KEY
A	0	1973-06-01	1952-01-01	1.00	M	GvTO7ugiWq...	Lxx3RjP90N...	NWBvf2eXcF...	xzH7hnk2G/...	covid19_deat...

The relevant data items are highlighted below with discuss of potential opportunities and limitations.

- Country Coverage: The DDI provides information on deceased individuals in the United States and Canada. However, there is no filter to differentiate the two. According to Datavant, about 99.6% of the data is on deaths in the United States, whereas the rest are from Canada. Thus, the DDI is essentially a U.S. mortality database. However, researchers should be aware that data on deaths in Canada are not easily excluded.

- Date of death and date of birth: Datavant's Death Index masks the date of death (or birth) to the year and the month of death (or birth) to protect the privacy of the deceased individual. The DDI sets the day of death (or birth) to the first day of the month. Consequently, the age at death is overestimated if the birth date occurs after the death date in their respective calendar months and underestimated otherwise.

- Death Date Imputation Flag: When this flag takes on a value of 1 (TRUE), the death date was imputed from the NTIS Social Security Administration's Death Master File (SSADMF) and could not be filled in by other sources. In these cases, the DDI inherits potential inaccuracies in the SSADMF until they are corrected. For imputed death dates, the day is set to the first day of the month. There are no cases in the database where dates of birth are imputed.

- Gender and Gender Probability: The data sources of the DDI do not provide gender information. The DDI creates two records for the same person, a male version and a female version. For each version, the DDI assigns a probability that the individual is the indicated gender by analyzing the individual's first name.

  The ambiguities in gender determination present challenges for analysis. The analysis of the first name may not be adequate to accurately determine the gender probability. Moreover, one would underestimate death counts by gender if one excluded records with gender probability around or equal to 0.5.

- Encrypted tokens: The DDI encrypts personal information into tokens to protect the individual's privacy.

  The following four tokens are available for all records in the DDI. These are probability tokens.
  - Token 1: Last Name + First Initial of First Name + DOB + Gender
  - Token 2: Soundex (Last name) + Soundex (First name) + DOB + Gender
  - Token 4: Last name + First name + DOB + Gender
  - Token 7: Last name + First 3 Characters of First name + DOB + Gender

  The following two tokens are deterministic tokens since they include a person's social security number (SSN). However, they are only available for individuals on the SSADMF and not available for all records in the DDI.
  - Token 5: SSN + DOB + gender
  - Token 16: SSN + First Name

  These tokens are useful to eliminate duplicate individuals in a dataset, and to match individuals across datasets.

- Token key: Each individual record has a token key and allows Personally Identifiable Information (PII) to be tokenized by Datavant and create a generalized Datavant token. The generalized Datavant token is consistent across all database owned by Datavant, such as healthcare data. This feature allows researchers to match individuals across databases and conduct additional analysis within the research database environment.

- Weekly updates: The frequent updates help to keep the DDI up to date by inclusion of recent deaths. For example, the DDI captures 90% of deaths (that will be eventually included) in a calendar month by the 14th of the following month. However, researchers should be aware that the weekly update file can also contain updates of old data for previously available records.

1.2 DATA RELIABILITY

In this section, individual records in the DDI are aggregated into death counts across years and age at death. These death counts are then compared to those from the HMD.

1.2.1 RECORD SELECTION

A combination of filters and tokens are used to select DDI records in order to appropriately aggregate death counts.

- Age at death: The DDI shows the date of death (or birth) as the first day of the death (or birth) month. Age at death is estimated by dividing the total number of months lived by 12.

- Gender: Records with gender probability less than or equal to 0.5 are excluded and include the otherwise identical records with gender probability greater than 0.5. Records with extreme gender ambiguity, those with gender probability equal to 0.5, only represent about 0.35% of the data in each year from 2001 to the present.

- Death Date Imputation: Records with Death Date Imputation Flag equal to one are excluded, to avoid dates of death that cannot be validated by other sources. Since 2017, almost all dates of death in the DDI can be validated by other sources.

- Token: Token 4 (Last name + First name + DOB + Gender) are used to select unique individuals. As described in Section 1.1, token 4 contains the most personal information among tokens that are available for all records in the DDI.

This record selection process allows a more accurate dataset to be obtained on the individual record level, but underestimates the total death counts by a small amount (for example, by excluding records with ambiguous gender). In Section 2 of this report, methods for estimating the population death counts are investigated by modeling the estimated capture rate by the DDI.

1.2.2 DEATH COUNTS BY YEAR

Table 1 compares the DDI-based annual death counts to those from the HMD up to 2018 which is the most recent year of currently available death data. While the DDI captures the majority of deaths, it still misses a meaningful portion. Moreover, the DDI missed the total death counts by a wide margin for years between 2006 and 2009.

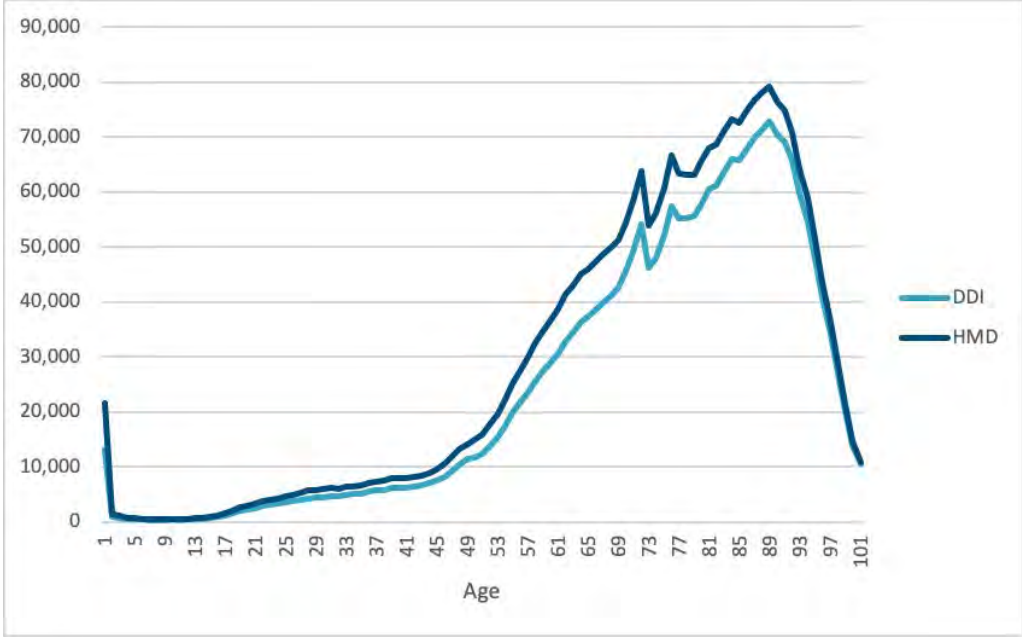
Table 1
TOTAL DEATH COUNTS

Year	DDI	HMD	Gender-Weighted DDI/HMD	DDI/HMD
2001	2,261,512	2,416,425	93.59%	93.59%
2002	2,297,897	2,443,387	94.05%	94.05%
2003	2,304,258	2,448,288	94.12%	94.12%
2004	2,237,573	2,397,615	93.33%	93.32%
2005	2,178,607	2,448,018	89.00%	88.99%
2006	1,959,439	2,426,264	80.76%	80.76%
2007	1,717,684	2,423,712	70.88%	70.87%
2008	1,506,704	2,471,984	60.97%	60.95%
2009	1,406,465	2,437,163	57.72%	57.71%
2010	2,182,160	2,468,435	88.41%	88.40%
2011	2,184,099	2,515,458	86.83%	86.83%
2012	2,034,500	2,543,279	80.00%	80.00%
2013	1,994,209	2,596,993	76.79%	76.79%
2014	1,962,143	2,626,417	74.71%	74.71%
2015	2,103,938	2,712,630	77.57%	77.56%
2016	2,049,033	2,744,248	74.67%	74.67%
2017	2,384,408	2,813,503	84.75%	84.75%
2018	2,447,668	2,839,205	86.21%	86.21%

1.2.3 DEATH COUNTS BY AGE

Figure 2 shows death counts by age based on the DDI and the HMD for 2018. Figure 3 shows the capture rate, defined as the ratio of the DDI death counts to the HMD death counts, by age in the year 2018. The DDI generally captures the profile of death counts (Figure 2), but unevenly across ages (Figure 3). Appendix A provides the annual levels and ratios by age and gender since 2013.

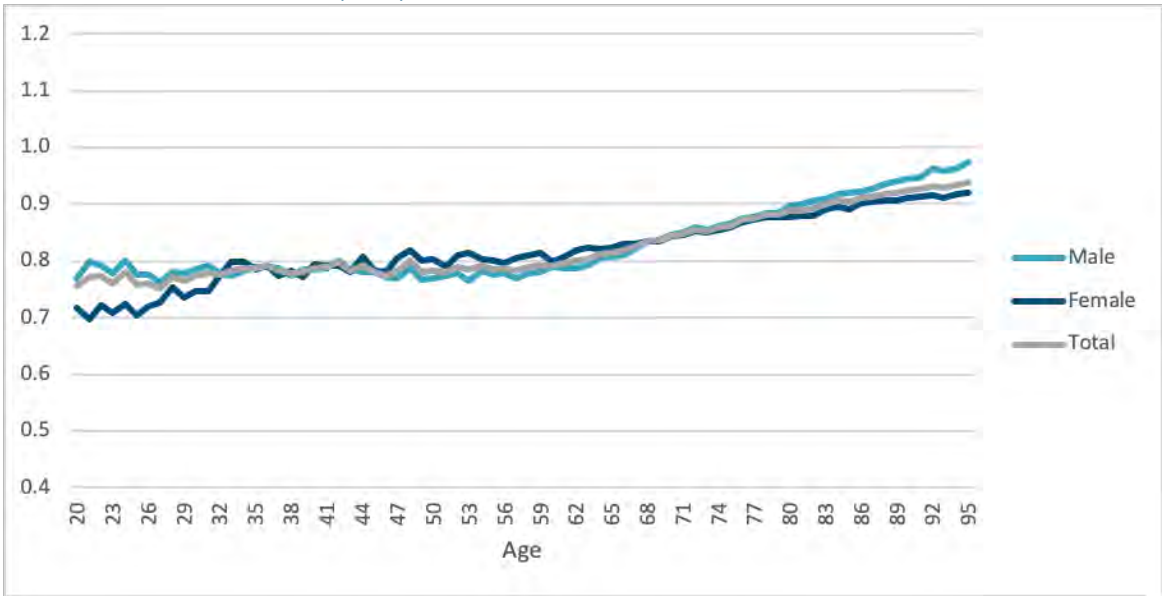
Figure 2
TOTAL DEATH COUNTS BY AGE (2018)



Deaths are concentrated in the 60-90 age interval, a pattern that is consistent in recent years, in the DDI and the HMD. For ages below 60, the capture rates tend to be in a stable range. Above age 60, the capture rate generally increases with age. Appendix A suggests that the capture rates have been improving across age since 2013, consistent with the aggregate data shown in Table 1.

While the DDI under-captures deaths in general, population death counts may be estimated by scaling the DDI death counts by its estimated capture rate. By modeling the DDI's capture rate, researchers can incorporate information in the DDI that is not concurrently available in other databases.

Figure 3
DEATH COUNT CAPTURE RATE (2018)



Section 2: Incorporating Datavant's Death Index in Mortality Analysis

In this section, ways to incorporate Datavant's Death Index in mortality analysis are explored. As an example, the SOA's RPEC_2014 mortality improvement model is adopted with the same assumptions in the SOA's MP-2019 report for making mortality improvement projections.

The annual update of the RPEC mortality improvement model is usually fitted with mortality data that lagged by two years. For example, the MP-2019 update of the RPEC model (which was published in October 2019) used mortality data up to 2017. By incorporating the Datavant's Death Index, the model with data up to 2019 (or 2020) can be fit to make projections and update previous projections.

2.1 MORTALITY IMPROVEMENT MODEL

The SOA's RPEC mortality improvement model uses a set of 1x1 mortality rates as input. The model calculates historical mortality improvement rates based on the smoothed natural logarithm of raw mortality rates from two-dimensional Whittaker-Henderson graduation of order 3. The model implements a two-year step-back from the most recent year of mortality data to address the edge effect before making projections. The first year of projection is the year before the most recent year of mortality data. Projections are based the average of an age-based and a cohort-based cubic spline satisfying a set of boundary conditions.

To project mortality improvement rates, the two-year step back is also implemented from the most recent year of mortality data. The same assumptions used in developing the SOA's Scale MP 2019 are adopted; specifically:

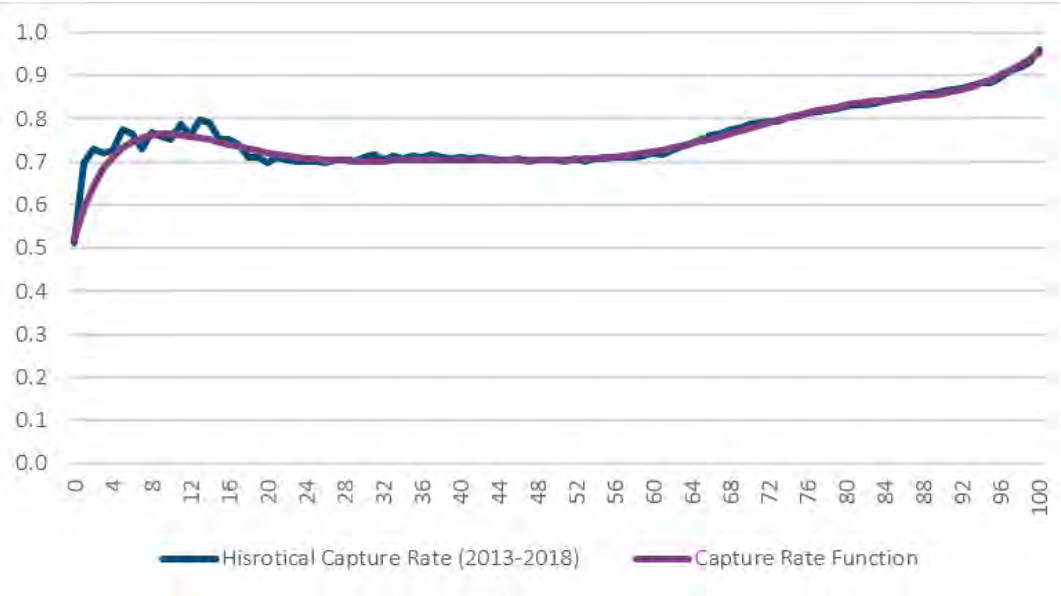
- Long term rate of improvement: 1.0% flat rate to age 85, decreasing linearly to 0.85% at age 95, and then decreasing linearly to 0.0% at age 115.
- Initial slope: 0
- Convergence period along fixed ages: 10 years
- Convergence period along fixed year-of-birth cohorts: 20 years

2.2 MORTALITY RATES WITH DDI DEATH COUNTS

To fit the mortality improvement model, mortality rates from the SOA up to 2017 are used, HMD for 2018 and develop mortality rates for 2019 and 2020 by incorporating the Datavant's Death Index.

Since the DDI under-captures deaths, the DDI death counts must be adjusted by their estimated capture rates. The capture rate function is the ratio of the DDI death counts to the HMD death counts by age. Historical death counts from 2013 to 2018 are used to fit a polynomial logistic regression. An examination of goodness of fit statistics (such as the AIC) suggests that the optimal orders of the polynomials are 7 for male and 4 for female. Figure 4 shows the average capture rates from 2013 to 2018 and the capture rate function.

Figure 4
CAPTURE RATE FUNCTION BY ATTAINED AGE (MALE)



To estimate raw mortality rates for 2019, the death counts from the DDI are first adjusted by the capture rate function, then divide the capture-rate-adjusted death counts by the population counts from the HMD.

Death counts for 2020 are estimated by first adjusting the first half-year death counts from the DDI by the capture rate function and then multiply the capture-rate-adjusted death counts for the first half by a subjective multiplier of 1.96 for the full year estimate. At the time of writing of analysis, data and analysis were available with completeness through June 2020. This subjective multiplier reflects a scenario that COVID-19 related mortality in the U.S. moderates in the second half. Population for 2020 are estimated as the 2019 population scaled by the average rate of increase in population from 2015 to 2019. The death count estimates are divided by the population estimates to estimate raw mortality rates for 2020.

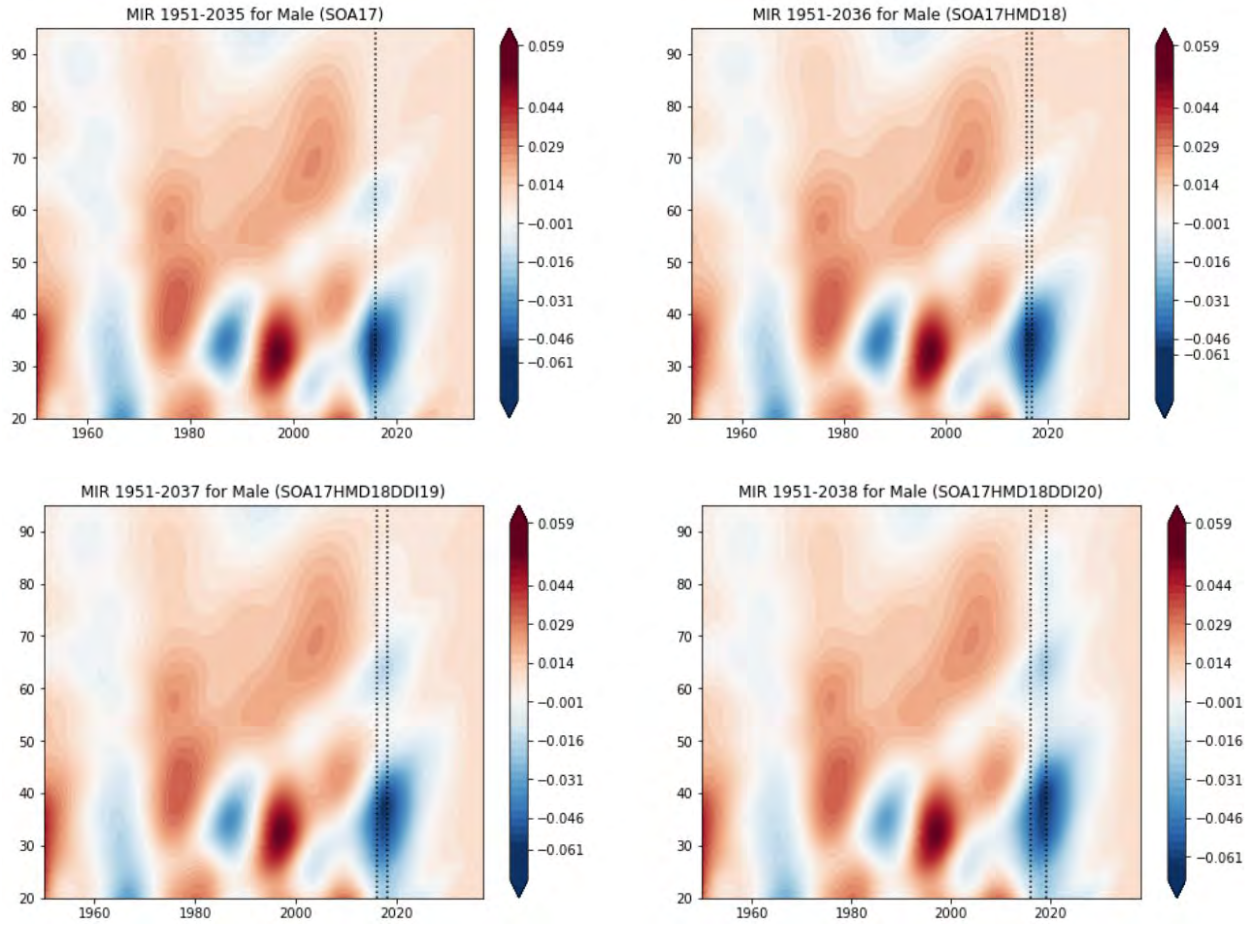
2.3 MORTALITY IMPROVEMENT RATES

Figure 5 shows historical and projected mortality improvement rates for male in a series of heatmaps. Each heat map represents the output of the mortality improvement model using a particular set of input data:

- SOA17: Raw mortality rates from the SOA up to 2017 are used as input to the model. 2017 coincides with the most recent year of raw mortality rates used in the MP-2019 report.
- SOA17HMD18: Raw mortality rates from the SOA up to 2017 and from HMD for 2018 are used as input to the model.
- SOA17HMD18DDI19: Raw mortality rates from the SOA up to 2017 and from HMD for 2018, plus estimated mortality rates from DDI for 2019 are used as input to the model.
- SOA17HMD18DDI20: Raw mortality rates from the SOA up to 2017 and from HMD for 2018, plus estimated mortality rates from DDI for 2019 and 2020 are used as input to the model.

In Figure 5 the first vertical line is at 2016, the first year of projection by the MP-2019 report. The second vertical line, if applicable, is the first year of projection, which is the year before the most recent year of mortality data used as input. This series of heatmaps show how estimated mortality improvement rates might change with extra years of information, one additional year at a time. For example, Figure 5 reveals that the deterioration in mortality improvement for male in the age ranges of 30-40 and 60-65 has persisted in the past few years.

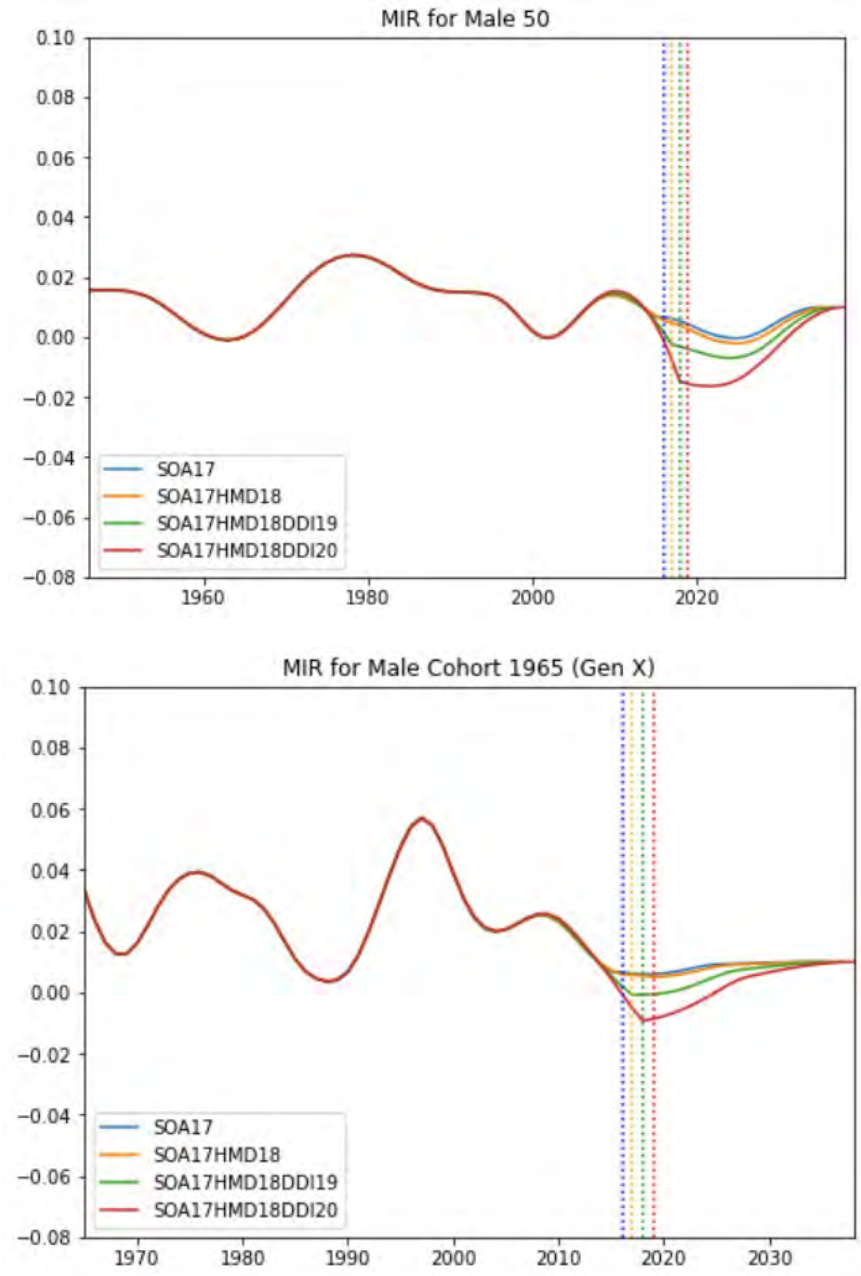
Figure 5
HEAT MAPS OF MORTALITY IMPROVEMENT RATES



Each inclusion of an extra year of mortality data also allows the updating of previous projections. For example, Figure 6 shows the time series plot of historical and projected mortality improvement rates for male age 50 and those born in 1965. Inclusion of recent mortality data one year at a time results in successively lower projected mortality improvement rates, especially for the near term. The relative gaps between the lines indicate the incremental impact of each year's data. The gaps between the green and orange lines and that between the red and green lines are relatively large. This suggests that 2019 and 2020 data are the main driver of reduced near term mortality improvement expectations.

Figure 6

MORTALITY IMPROVEMENT RATES (MIR) FOR SELECTED AGE/GENDER/COHORT COMBINATIONS



Appendix B provides the historical and projected mortality improvement rates by gender for select ages and year-of-birth cohorts.

Section 3: Acknowledgments

Thanks to the Society of Actuaries and the Actuarial Science program at Columbia University for their generous support. Our gratitude goes to the SOA for reviewing our presentation and discussions, and to Thomas J. Murphy and Ira Kastrinsky at Columbia University for their insights and support. Thanks to the COVID-19 Research Database partners, who are acknowledged at https://covid19researchdatabase.org, for generously supplying the data.

Appendix A: Death Counts from Datavant's Death Index

Figure A1

DEATH COUNTS BY AGE (MALE)

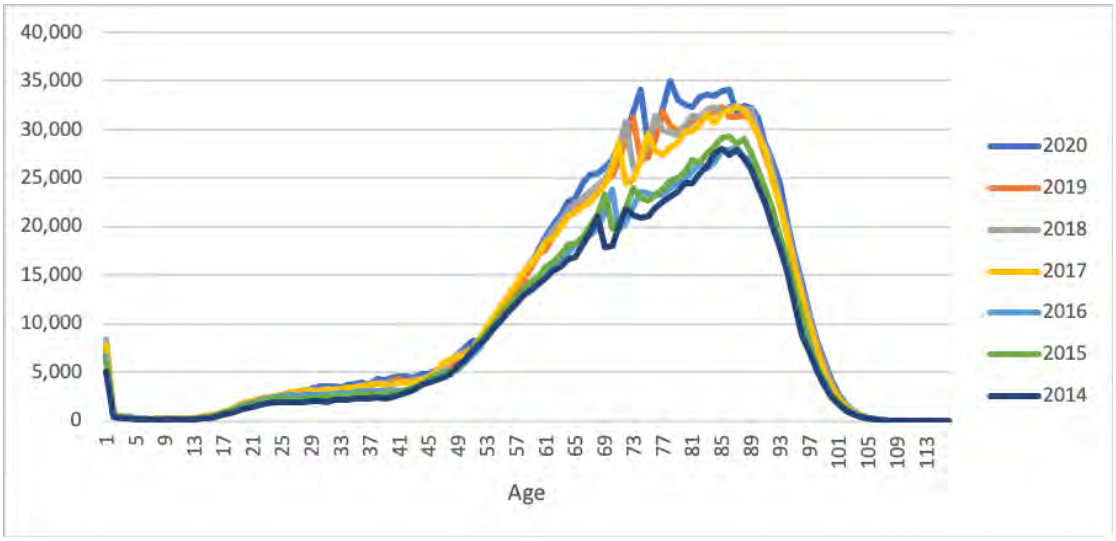


Figure A2

DEATH COUNTS BY AGE (FEMALE)

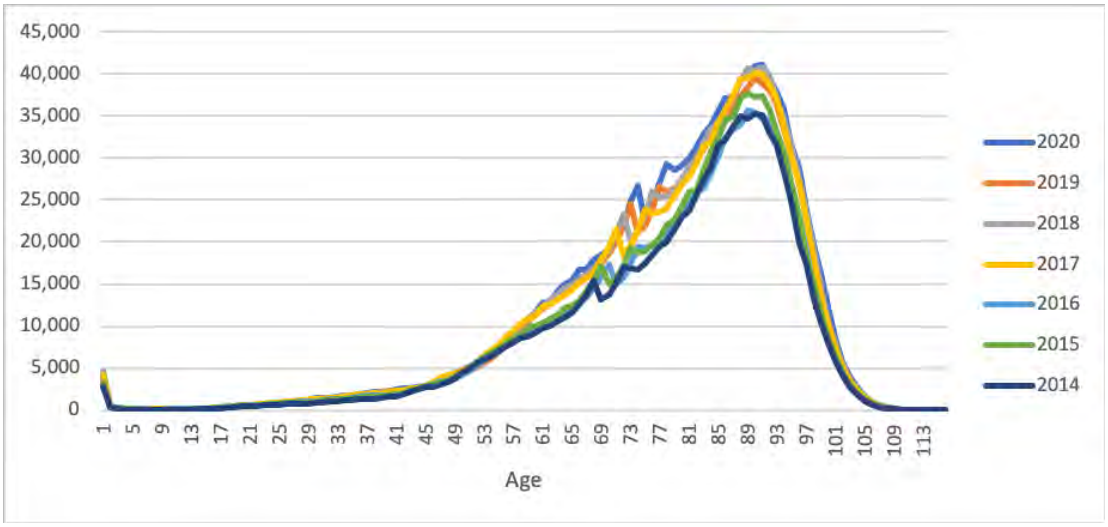


Figure A3
DEATH COUNT CAPTURE RATES (MALE)

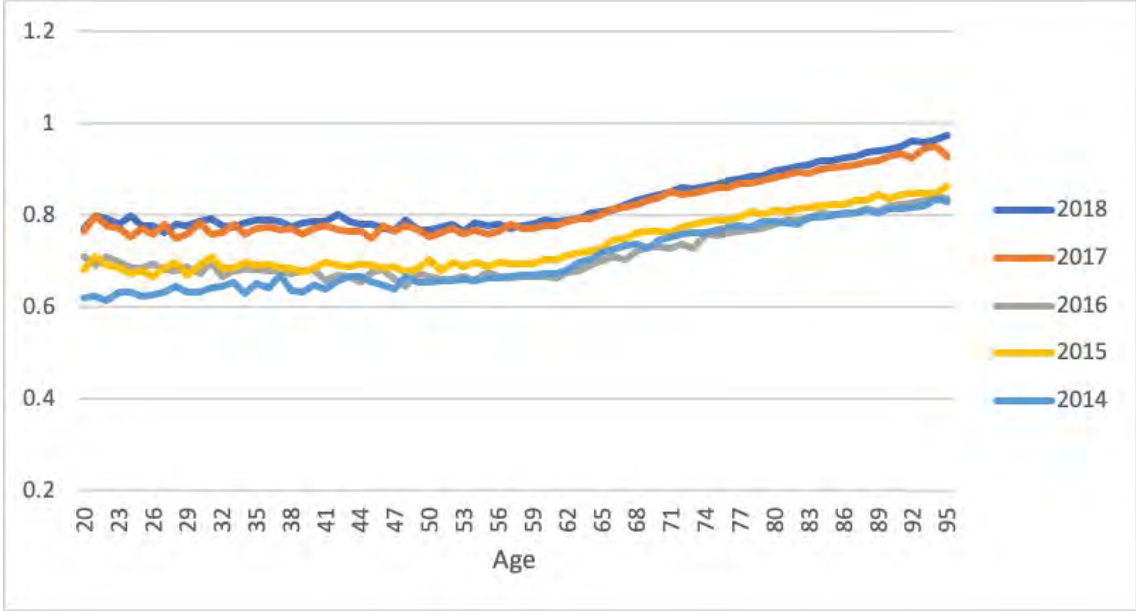
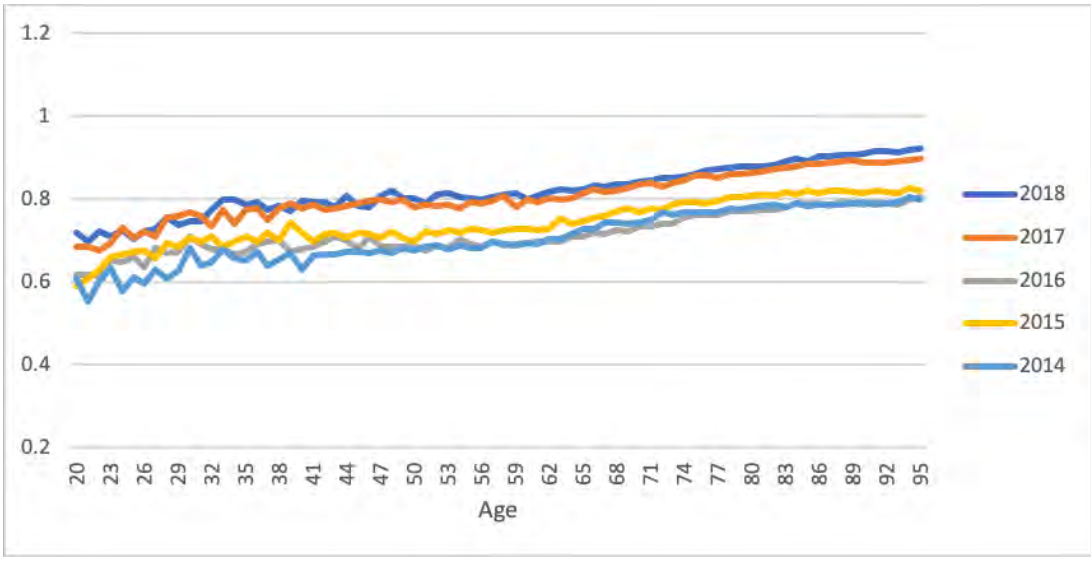


Figure A4
DEATH COUNT CAPTURE RATES (FEMALE)



Appendix B: Mortality Improvement Rate Updates

Figure B1

HISTORICAL AND PROJECTED MORTALITY IMPROVEMENT RATES (MALE, SELECT AGES)

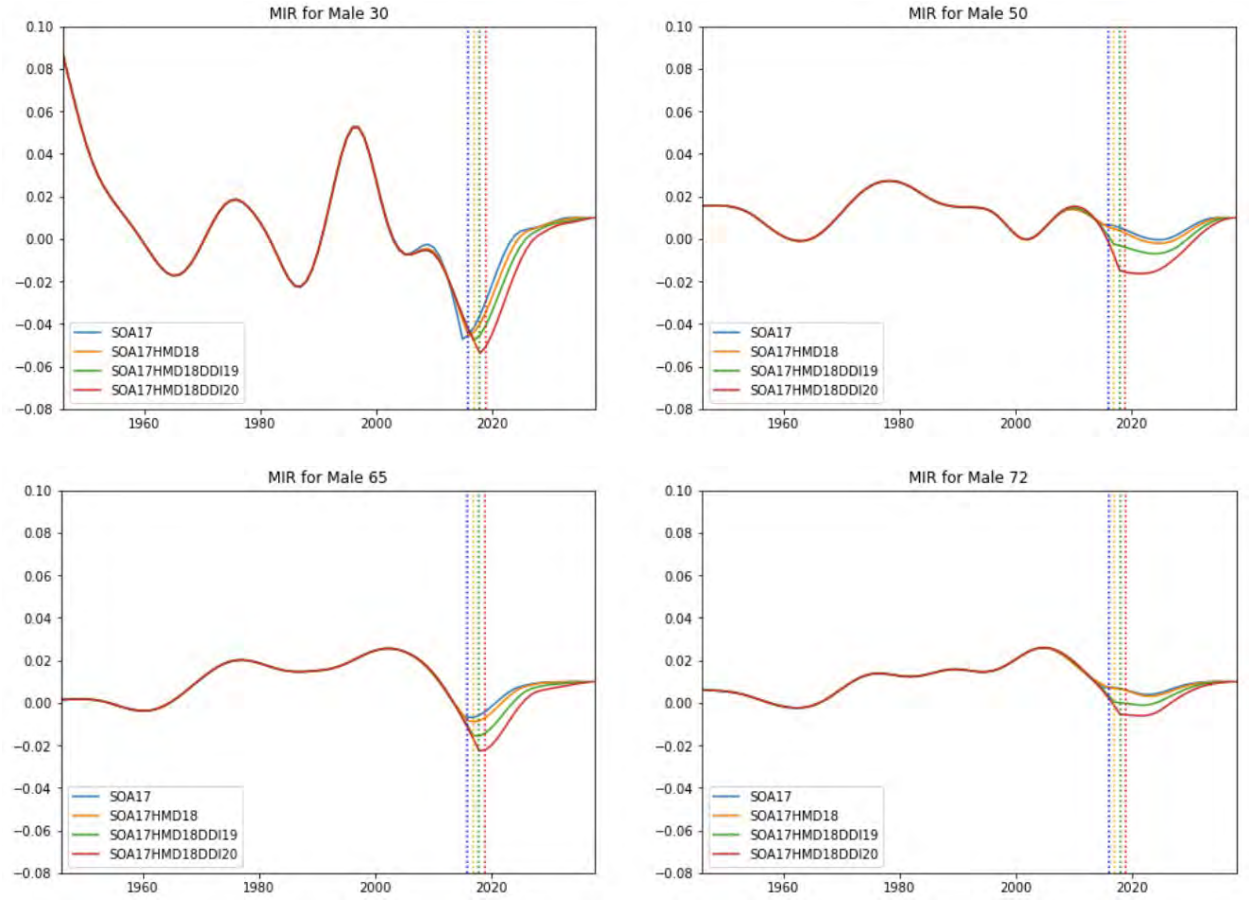


Figure B2

HISTORICAL AND PROJECTED MORTALITY IMPROVEMENT RATES (FEMALE, SELECT AGES)

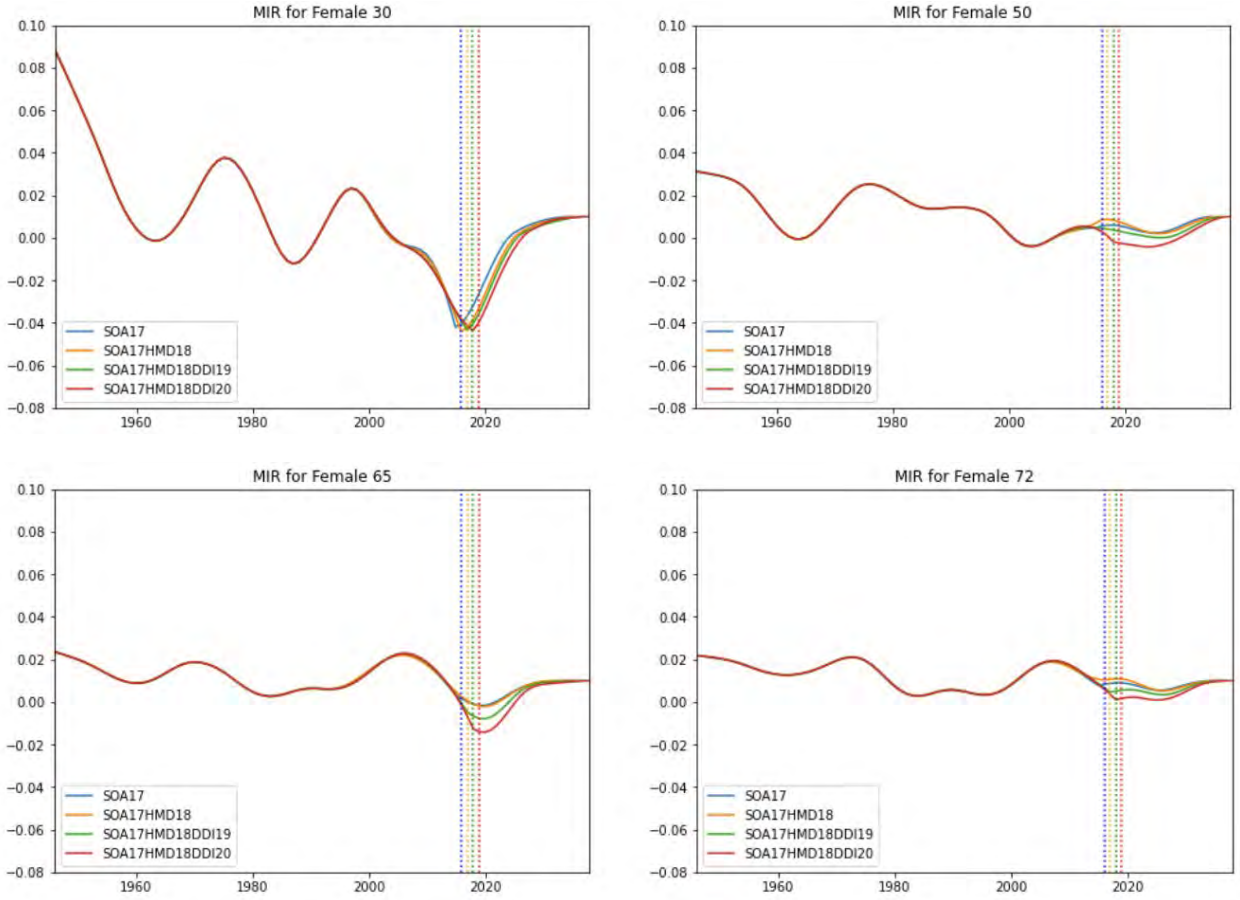


Figure B3

HISTORICAL AND PROJECTED MORTALITY IMPROVEMENT RATES (MALE, SELECT YEAR-OF-BIRTH COHORTS)

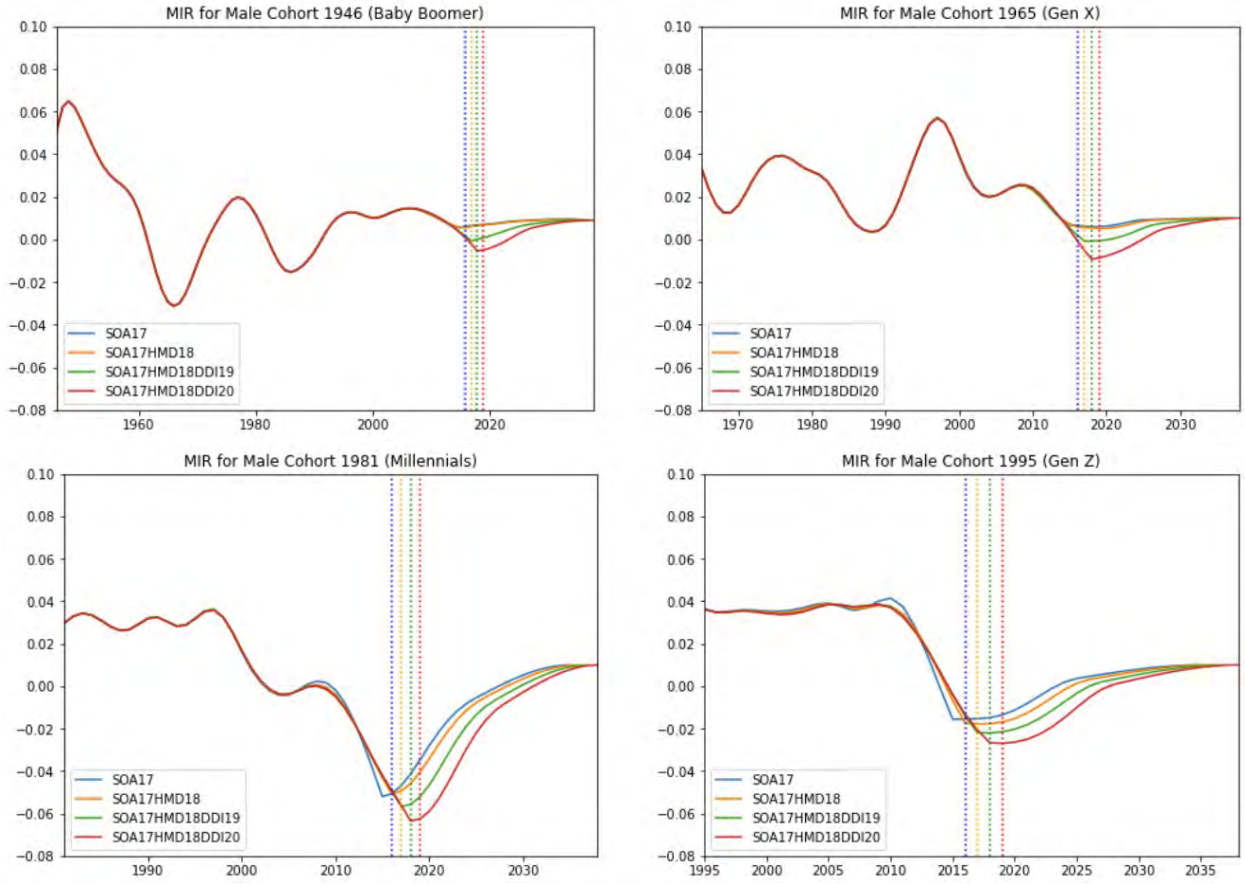
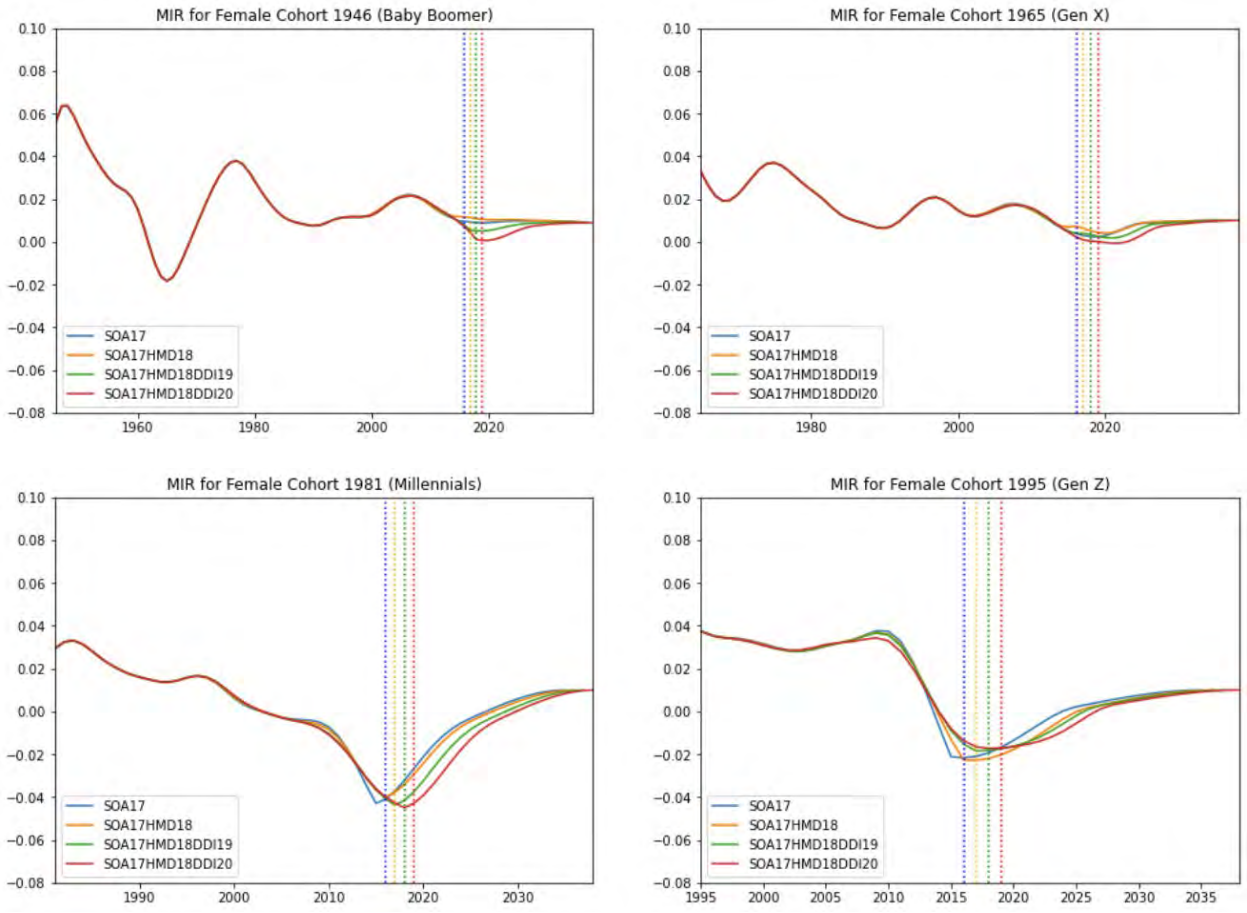


Figure B4

HISTORICAL AND PROJECTED MORTALITY IMPROVEMENT RATES (FEMALE, SELECT YEAR-OF-BIRTH COHORTS)



References

COVID-19 Research Database. https://covid19researchdatabase.org (Data Downloaded as of 2020-09-26).

Hardy, Mary R., Long Term Actuarial Mathematics Supplementary Note, Society of Actuaries 2017.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (Data Downloaded of 2020-09-26).

Knorr, Frank E., Multidimensional Whittaker-Henderson Graduation, Transactions of Society of Actuaries 1984 Vol. 36, p213-255.

Macaulay, Frederick R., The Smoothing of Time Series, National Bureau of Economic Research 1931.

Society of Actuaries. 2014. Mortality Improvement Scale MP-2014 Report. Schaumburg: Society of Actuaries. https://www.soa.org/globalassets/assets/files/research/exp-study/research-2014-mp-report.pdf

Society of Actuaries. 2019. Mortality Improvement Scale MP-2019 https://www.soa.org/globalassets/assets/files/resources/experience-studies/2019/mortality-improvement-scale-mp-2019.pdf

About the Society of Actuaries

With roots dating back to 1889, the Society of Actuaries (SOA) is the world's largest actuarial professional organizations with more than 31,000 members. Through research and education, the SOA's mission is to advance actuarial knowledge and to enhance the ability of actuaries to provide expert advice and relevant solutions for financial, business and societal challenges. The SOA's vision is for actuaries to be the leading professionals in the measurement and management of risk.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policymakers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public policymakers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement and other topics. The SOA's research is intended to aid the work of policymakers and regulators and follow certain core principles:

Objectivity: The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

Quality: The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and nonactuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

Relevance: The SOA provides timely research on public policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

Quantification: The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

Society of Actuaries
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org