Calibrating ZIP+4 VitaCurves

Our "Zooming in on ZIP codes" paper introduced our VitaCurves, a series of mortality tables derived from pooling pension plan data which enable plans to use a baseline longevity assumption tailored to the true diversity of their participants.

In this paper we provide an overview of the techniques used to build these tables. These techniques mirror those well-established for analyzing UK and Canadian pension plan longevity, which have been peer reviewed by actuarial bodies in both of those countries[1,2]. The paper is aimed at a technical actuarial audience interested in understanding more about the methods we use and should be read in conjunction with "Zooming in on ZIP codes" and our paper summarizing the data used. Further detailed technical information can be provided on request.

Throughout this paper, we refer to "ZIP" and "ZIP+4" codes as the 5-digit and 9-digit versions of "zone improvement plan" codes used for postal delivery in the United States, respectively.

If any of the technical terms in this paper are unfamiliar to you, you can find definitions and related terms in our online Lexicon of Longevity. Just visit www.clubvita.net/glossary and search for a term.

1 An overview of our approach

At the heart of our approach is a statistical method known as "Generalized Linear Modeling" (GLM) which links mortality rates to the values taken by a range of factors known to be relevant to the chance of dying (for example an individual's age, income and ZIP+4 information).

In terms of conventional actuarial notation, we can think of the chance of someone dying (typically denoted q_x) as being determined not just by their age x but a whole host of potentially relevant factors:

$$logit(q_x) = a\ linear\ function\ of\ predictors\ (e.g.age, income, ZIP+4)$$

Where $logit(q_x)$ is the logistic transformation i.e.

$$logit(q_x) = \ln\left(\frac{q_x}{1-q_x}\right)$$

The logistic transform is an evolution of the basic log transformation proposed by Gompertz. In the world of Generalized Linear Modeling it is also the "natural" scale for analyzing events like the numbers dying at each age which are expected to follow a Binomial distribution.

Our approach then fits a series of these models to the data, with each model optimized to provide a good fit to the underlying data. In order to fit these models, we need to identify which of the potential longevity predictors in our data is relevant, and how best to include these – we summarize this in **Section 2**, with **Section 3** covering how we create our ZIP+4 groups and annuity bands for use in the model. **Section 4** describes how we fit the model to the data, with **Section 5** describing the range of detailed goodness of fit tests applied to the model and **Section 6** checks done for internal consistency of the model. We conclude in **Section 7** by describing how we extend the fitted model to provide

[1] For the UK see: "What longevity predictors should be allowed for when valuing pension scheme liabilities"

[2] For Canada see "Key Factors for Explaining Differences in Pensioner Baseline Mortality"

VitaCurves which cover the full age range (i.e. to include younger and older ages outside the range of the data fitted to).

2 Allowing for different longevity predictors

All our VitaCurves use the GLM construct, with each reflecting a specific combination of data fields identified as being "predictive" of longevity (known as "rating factors"). Potential predictors available within our initial dataset for each plan participant include:

- Age
- Gender
- Status (healthy annuitant, disabled retiree or surviving beneficiary of a deceased retiree)
- Type of annuity (single or joint life)
- Annuity amount in payment
- Socio-demographic and geographical data identified from the participants ZIP or ZIP+4 code
- Collar type of (former) employment
- Industry of (former) employment[3]

Traditionally, actuaries have fitted ("graduated") life tables by first segmenting data into groups with like values of a predictor and then smoothing to this data. This is the approach used for instance by the Society of Actuaries in their Private Sector Pension Plan study (i.e. RP14 and Pri-2012 tables) and their Public Sector Pension Plan study (Pub-2010 tables). This helps to reduce heterogeneity (variability) in the underlying data, but at the expense of reducing the data volumes in each group, in turn reducing the certainty in the mortality rates and/or the potential number of groups that can be formed. The application of the GLM techniques enables a wide range of internally consistent tables to be fitted simultaneously across a range of variables. This makes maximum use of the available data, improving confidence in the resulting tables while creating a model that captures the diversity of the underlying population.

It is generally preferable wherever practical to use a predictor directly in the GLM (known technically as treating the variable as a "**covariate**"). Different values of the rating factor are then modeled simultaneously within the GLM, maximizing the amount of data used to infer the impact of any particular rating factor. However, for some predictors it remains necessary to segment the data into groups before fitting the GLM to each group. Treating a longevity predictor in this way (as a "**stratifier**") is generally preferred when:

- The *meaning* of other variables differs across the predictor, greatly complicating the fitting of a GLM; or
- The *shape* of mortality with age is very different depending on the value of that variable; or
- The *data* spans different age ranges for the different values of that variable

[3] Specifically, one of 7 categories: Auto Industrial Goods and Transportation; Basic Materials; Banking, Financial Services and Insurance; Chemicals, Oil, Gas & Utilities; Consumer Goods and Food/Drink; Healthcare and Hospitals; Other

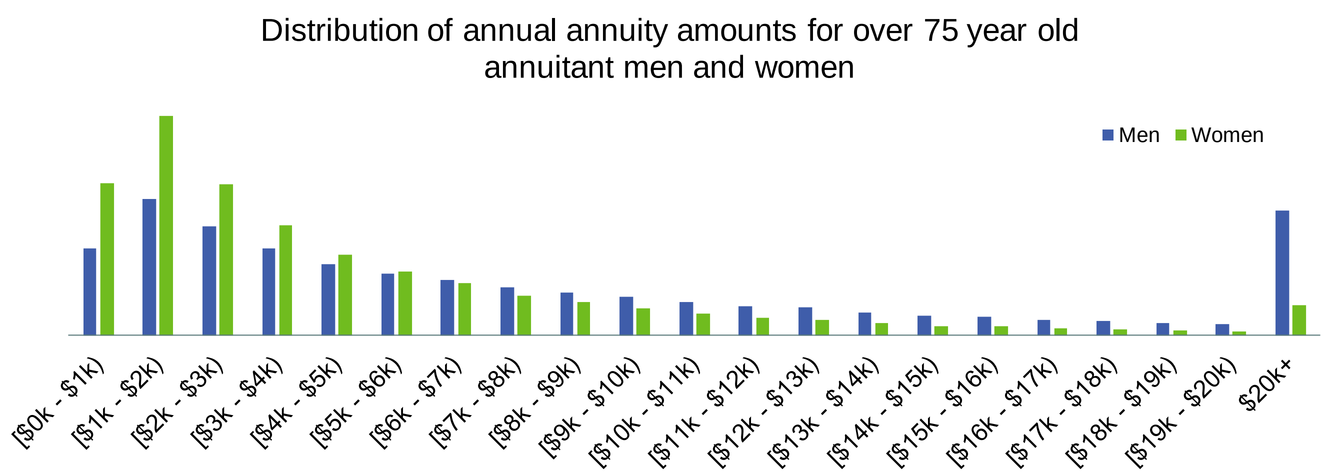
2.1 Splitting the data into sensible groups for modeling (stratification)

We split the data into six groups determined by participants' gender and status (healthy annuitant, disabled retiree or surviving beneficiary of a deceased retiree), for the reasons set out below.

Why split men and women?

In the US, like in Canada and the UK, the relative importance of different covariates can vary between the genders. For example, we find that the annuity income of a woman is often less informative to their longevity than it is for a man. In many territories this is a legacy of both historical working patterns, and the history of pension plan participation of men and women, which has led to women having accrued lower pensions than men among older generations. In turn, the income of this generation of women is often determined more at the household than the individual level. Consequently, we prefer to split the data between men and women to ensure we correctly capture the importance of different rating factors for men and women.

Figure 1:



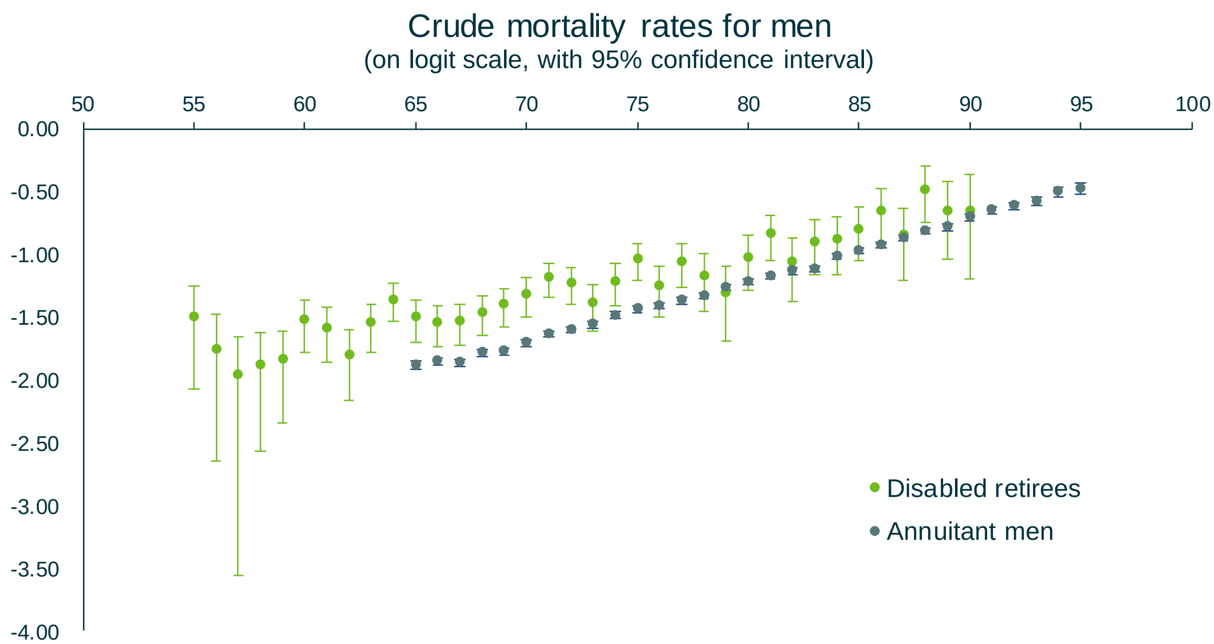
Note: Chart includes all annuitants (exc. disabled retirees) aged over 75 as of January 1, 2019 who were exposed to risk during the period 2014-2016 and so potentially were included in our modeling.

Healthy annuitants and disabled retirees

Disabled retirees tend to retire younger than healthy annuitants – on average by 5½ years in our dataset – and tend to have curtailed life expectancy. We split out the data for disabled retirees to reflect the differences in the age range for which we have credible data. Also, the observed proportions dying at each age ("crude mortality rates") show a different shape at younger ages which we need to take care to ensure the model captures (see Figure 2).

As a relatively small proportion of the data relates to disabled retirees (just 9,095 lives or around 1.5% of plan participants) we do not attempt to identify the impact of other factors on the longevity of disabled retirees.

Figure 2:



Surviving beneficiaries of plan participants

We also split out the data relating to the surviving beneficiaries of deceased retirees. There are fewer individuals in this group (67,752 women and 16,053 men), limiting our ability to identify the impact of other factors on the longevity of joint life survivors. They also tend to be older and the meaning of some of the rating factors differs (for example the annuity amount is typically a reduced proportion of the original plan participant). For this group, our VitaCurves use annuity amount and ZIP+4[4].

The resulting rates for this group show levels of elevated mortality consistent with that seen in other studies[5], which we believe are broadly indicative of the mortality rates for these individuals, notwithstanding the limitations of the data reporting procedures for these lives[6].

2.2 Identifying key variables which influence mortality for each group (covariates)

Our work in other territories has identified that the following variables tend to be highly predictive of longevity:

- **Income** via annuity amount
- **Lifestyle** via ZIP+4 longevity group
- **Collar type** of former employment

For US pension plan participants, all three of these variables are clearly associated with variations in mortality, and so life expectancy, as shown in Figure 3. Analysis by the Society of Actuaries[7] on private pension plan data containing income, collar type and industry has also noted the strong variations by both income and collar type.

---

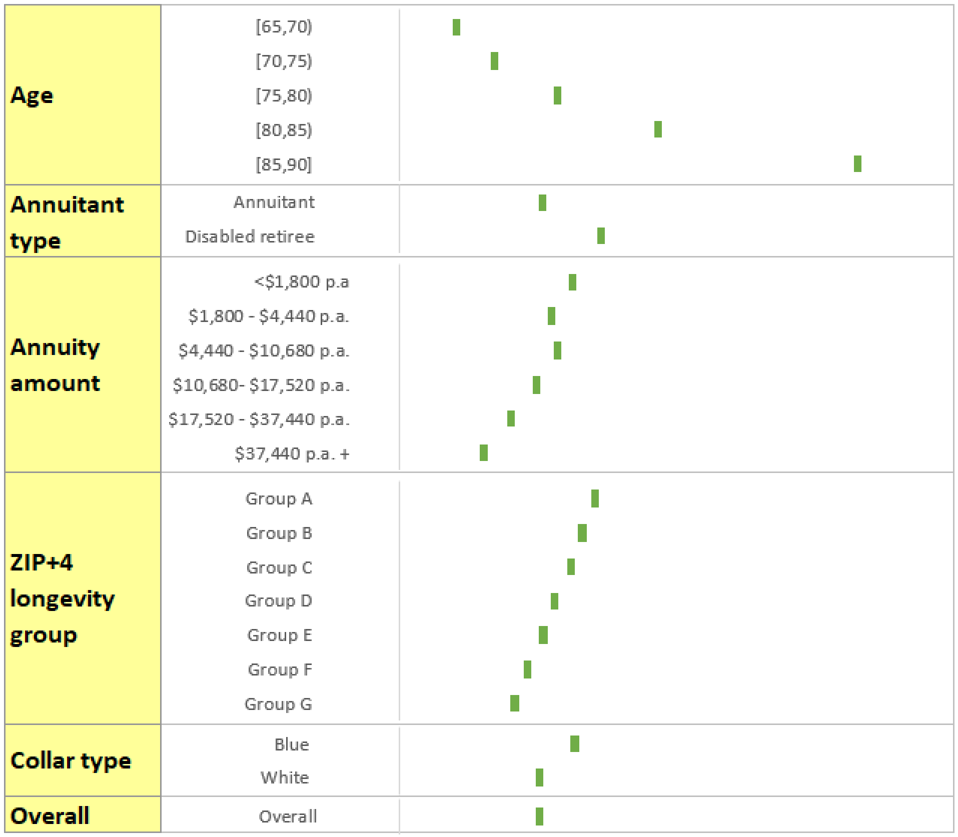
[4] We do not allow for collar type as this relates to the original plan participants employment and so less relevant to the surviving beneficiary

[5] See our publication "Grieving Widows" and Society of Actuaries' "Exposure Draft: Pri-2012 Private Retirement Plans Mortality Tables"

[6] See our paper on underlying data available at: https://www.clubvita.us/zooming-in-on-zip-codes

[7] See https://www.soa.org/globalassets/assets/files/resources/experience-studies/2019/pri-2012-mort-tables-exposure-draft.pdf

Figure 3:



NB: Crude mortality rates based on men aged 65-90 and after application of quality control criteria for all variables.

These predictors are not independent – often annuity amounts will be higher for white collar employees, and income is closely correlated with other socio-economic factors capture by ZIP+4. As such, we need to verify whether all three predictors are relevant to include in our modeling. We do this by sequentially adding potential rating factors to the model with the aim of improving fit. This “stepwise method” highlights the benefit of adding specific factors to the model, and combinations thereof, but also the point where we gain limited benefit from adding further predictors.

Figure 4 (next page) illustrates the results of this for healthy annuitant men. The longer the bars extend the better the model “fits” to the data[8]. We can see how:

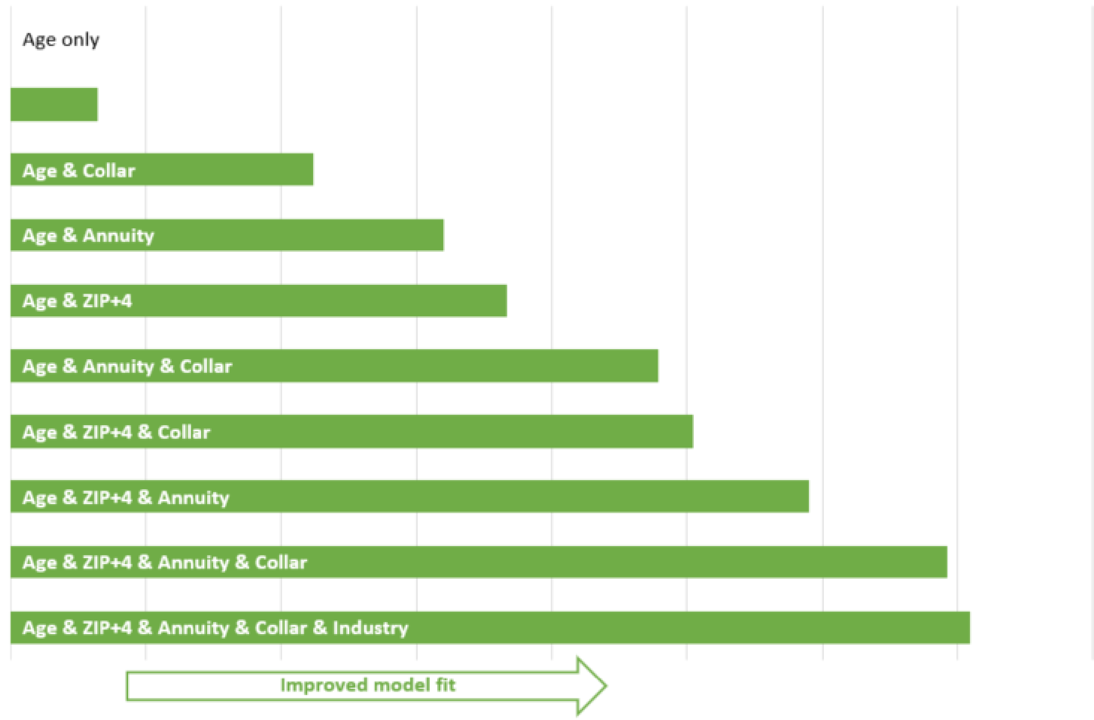
- Allowing for any one of income, lifestyle (ZIP+4), collar type and industry improves the model fit, with ZIP+4 being the single best factor *(2nd to 5th bars*)
- Using two or more factors materially improves the fit of the model, with the best pair of variables ZIP+4 and annuity amount *(6th to 8th bars*)
- A model which includes income, lifestyle (ZIP+4) and collar type is preferable to a model which only allows for two of these variables *(9th bar compared to 6th to 8th bars*)

We can also see how adding industry as an ***additional*** variable adds limited additional insight when income, ZIP+4 and collar type are available *(10th bar compared to 9th bar).* This is consistent with our work in the UK and Canada,

[8] Technical Note: The bars represent the improvement in Akaike Information Criteria upon adding additional variables

which has suggested that variations in mortality between industry (and indeed between public and private sector plans) can be largely explained by differences in the mix of incomes, lifestyles (as proxied by ZIP+4) and collar type[9]. Given this limited additional insight, and how heterogeneous some industries are, we have elected not to include industry at the stage, although we will continue to investigate the benefit of industry and other possible predictors as we grow our data set.

Figure 4: Improvement in fit of model by adding different potential predictors



3 Forming "buckets" of like individuals

The information we have collected contains large number of potential covariate values and some grouping is needed before we carry out the fitting. For example:

- **ZIP+4 information:** There are around 46 million ZIP+4 codes, covering over 130 million residential properties (and over 5 million commercial buildings). Even with the use of a third-party provider to convert these ZIP+4 codes into 58 different types of neighborhoods, we still have too many points to credibly model each type of neighborhood as a separate value.
- **Annuity income:** The annuity can take a range of values, from a few dollars to many tens of thousands of dollars. In principle this could be captured using a continuous variable in the GLM, but this would lead to every individual annuitant in a plan having a unique VitaCurve. This would be unwieldy for most plan valuation systems. We have elected to use an alternative approach of initially grouping annuity amounts into percentage splits of the data (either 2% or 5% depending on data volumes) within each of the subgroups (gender and annuitant type).

[9] This is also consistent with the conclusions of the SoA's multivariate analysis [accompanying the draft Pri-2012 tables](), which encouraged any assessment of industry/plan specific effects to be based on plan-specific experience, if credible.

The parsimony principle

A key criterion in the development of a statistical model is parsimony;

> ***"a simpler model with few rather than many parameters is favored over comparatively complex ones, provided they fit the data about equally well."***

The number of parameters in a model can be reduced by either:

- reducing variable to a simpler form; or
- if possible, omitting the unnecessary variables in a model

For example, imagine the inclusion of a 50-level discrete covariate in a regression model describing the geographical location of an annuitant based upon which state they live in. If mortality experience can be described equally as well with the inclusion of a simpler geographical variable (Northeast, Midwest, South, West) then this model has materially fewer parameters (46 fewer) and is preferred over the more complex model.

For the purposes of assessing parsimony we need a measure of goodness of fit which penalizes for complexity. We generally use the Akaike Information Criterion which captures the goodness of fit via the (log-) likelihood (i.e. a measure of how likely the observed data is under the fitted model) with a penalty based upon the number of variables used (i.e. to avoid "over-fitting" to the data).

Having reduced ZIP+4 to 58 groups, and annuity income to 50 groups for retirees[10], we now have a more tractable problem – namely how to further group these values to ensure we have a model which adheres to the parsimony principle (i.e. sufficient groups to capture the major variations in longevity, without having an excessively complex model). We achieve this via a two-stage process:

1. Identify a potential 'optimal' grouping
2. Verify the grouping provides a practical solution to users (and if not restart the process)

3.1 Finding an 'optimal' grouping

Naively, we might consider an approach to determining the clustering which looks at all potential ways of grouping each of the covariates and fits the GLM for each. The combination of groupings for Zip+4 and annuity amount which results in the best fit would then be considered the most parsimonious and thus, optimal. The sheer number of possible groupings for ZIP+4 codes alone makes this impractical. For example, there are over $2\text{x}10^{45}$ possibilities of reducing the 58 neighborhood types into 7 groups which we ultimately identify as being optimal.

Fortunately, statistical techniques have been developed to help overcome this challenge. These are designed to identify very good candidate clusterings, while accepting any method is unlikely to be truly optimal. The approach we have used is known as recursive partitioning[11]. Under this approach, the variable is first split into two groups, then one of these groups is split into two, and so on until there is evidence of limited benefit of further splitting in terms of goodness of fit vs complexity of the model. At each stage, splits are identified based upon a minimum required improvement in how much of the variation within the data is explained at each increase in the number of groups. This assessment of when to cease splitting into further groups is based upon cross-validation, whereby we split the entire

[10] and fewer for surviving beneficiaries reflecting the lower data volumes

[11] In our work on similar datasets in the UK we have considered a number of possible approaches and noted they give rise to similar results

data into several equal subsets, fit the model to all the data excluding one of these subsets, and then test the performance of the model in predicting the outcomes for the subset of the data not used for fitting. This is done for each of the subsets in turn and an average prediction error calculated. Once the prediction error falls below a certain level (typically one standard error) we have entered the realm of "good models" and a point is usually reached where the prediction error starts to plateau (i.e. falls very little for each additional group added), indicating that there is limited benefit from creating further groups[12].

3.2 Ensuring a practical solution

It is possible that the clustering method can identify some very small groups which appear to have very different longevity, particularly for the less populous ZIP+4 neighborhood types. These groups are likely to be reflecting, at least in part, data volatility, rather than genuine differences in longevity. To ensure this is avoided for the ZIP+4 model we also:

- Require that a minimum amount of the data (2%) is included in each group during the partitioning process
- Identify any neighborhood types with very small data volumes and anomalously high/low life expectancies based upon the raw data.

  These types are likely to be exhibiting artifacts of noise in the data and so we pre-group these with a neighborhood type of similar socio-economic characteristics. This is done for a very small number of neighborhood types.

It is also possible for the clustering method to identify groups which contain large volumes of data yet have very modest (but statistically significant) differences in mortality rates. These differences may not be material when converted to life expectancies and annuity values. In such instances, retaining the extra groupings is likely to be spurious, particularly once further covariates are included in the GLM providing additional differentiation between individuals. We therefore seek that the crude life expectancies at age 65 based upon the raw mortality rates (i.e. prior to the model fitting) and curtate to the maximum age used in the fitting:

- differ by at least 6 months between consecutive groups; and
- are statistically significantly different at the 95% confidence level between consecutive groups

The first of these criteria provides confidence that we are capturing financially material differences, the second that there will continue to be material differences in life expectancies between groups once we allow for other covariates in the GLM.

Where either of these criteria are not met, we have the choice to either revisit the partitioning to identify an alternative grouping, to combine groups, or where the criteria are only just breached allow it through. Judgement is applied as to which approach is adopted based upon the results from a series of diagnostics.

3.3 Resulting groups

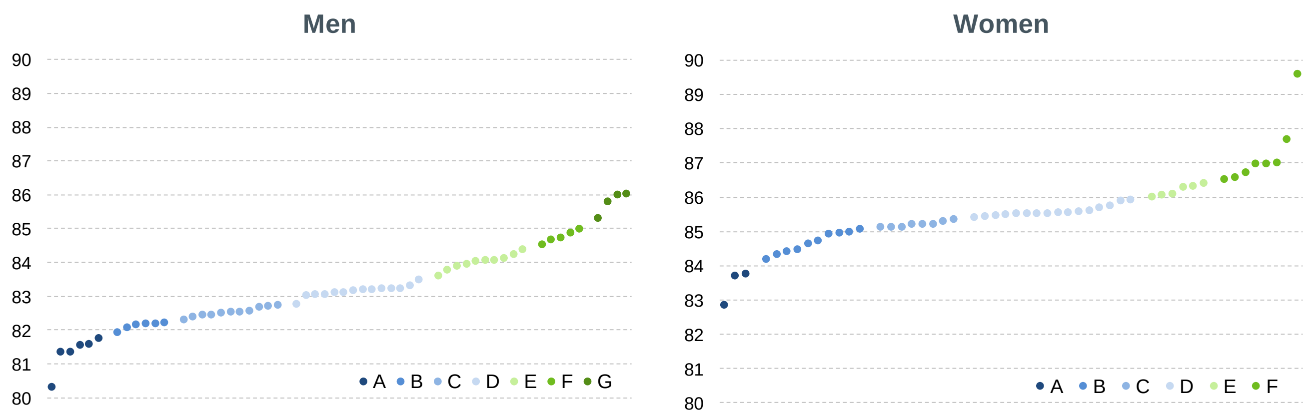
In the case of annuity amount this approach is adopted separately for each of the 4 groups healthy annuitant men/women and surviving beneficiary men/women leading to the following (annual) annuity income bandings:

[12] For those wishing to read more on the methods described here please see *Modern Applied Statistics with S* by Venables & Ripley (published by Springer, 2002)

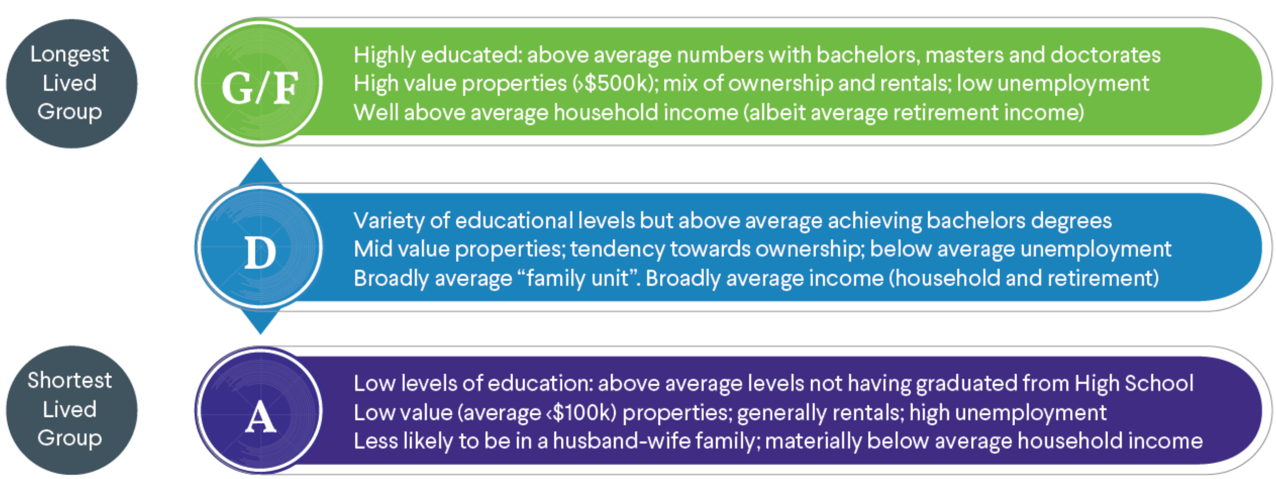
Band	Annutant men	Annuitant women	Surviving beneficiary men	Surviving beneficiary women
1	<$1,800	$1,680	<$3,000	<$2,400
2	$1,800 - $4,440	$1,680 - $9,120	$3,000+	$2,400 - $7,800
3	$4,440 - $10,680	$9,120+		$7,800+
4	$10,680- $17,520			
5	$17,520 - $37,440			
6	$37,440+			

For ZIP+4, the grouping is carried out separately for men and women (as socio-economic factors can influence men and women differently) using the healthy annuitant dataset and then, for reason of practicality, inherited by the surviving beneficiary curves. This results in 7 groups for men, labeled A (shortest lived socio-economic group) to G (longest lived socio-economic group). For women, 6 groups are identified (labeled A to F).

Figure 5: Crude life expectancy at age 65 split by marketing groups and colored by longevity group



The graphic below summarizes the key features of these groups.



4 Fitting the model

4.1 What ages and years?

In order to fit our model, we first need to decide which years we will fit to, and what age range the data can be used to reliably inform the model. Our data includes experience up to, and including, 2016.

We want the results of the longevity model to be as relevant to current baseline longevity as possible. The more recent the period fitted to, the less the need to apply adjustments for the passage of time. The volatility associated with good and bad winters, flu seasons (such as the high severity 2017-18 season[13]), etc. means it is also preferable for any model to incorporate multiple years, rather than focusing on the most recent year alone. We have chosen to base our modeling upon the most recent three calendar years of data (i.e. 2014-2016). This means that the resulting mortality rates are applicable[14] to a retiree of that exact age on January 1, 2015.

The age ranges we fit to are determined by data volumes. The sparser the data, the lower the certainty we can have about the underlying mortality rates at each age. We therefore restrict the fitting of the Generalized Linear Model for each group of data (annuitant type/gender) to the age ranges in the table to the right.

Group	Age range fitted to
Annuitant men	65 – 95
Annuitant women	65 – 95
Disabled retiree men	55 – 90
Disabled retiree women	60 – 80
Surviving beneficiaries, men	65 – 95
Surviving beneficiaries, women	60 – 95

4.2 What data?

The model is fitted to the data collected from pension plans. This data has been through pre-processing (including calculation of initial exposed to risk in each calendar year for an individual) and quality control by Mercer, and we also apply some additional controls to the data including:

- **Earliest usable dates:** Data only enters the analysis from the first date we are comfortable that we have complete recording of deaths.
- **Latest usable dates:** Where there is evidence of incurred but not reported deaths, we have restricted the period the plan can contribute for.
- **Quality flags:** We exclude plans from the data used to fit the model if there are concerns over the availability of data for that plan, or where any missing/poor quality data is biased between lives and deaths.

More information on the dataset, the quality controls applied to that data, and the volumes of data are contained in our accompanying paper *Data Underpinning ZIP+4 Curves*.

The data table below shows the life years of exposure, and the number of deaths contributing to our most-detailed VitaCurves for different groups. In each case, this relates to the final data used to fit the model (i.e. after application of the restriction to 2014-2016 calendar years and the age ranges described above).

---

[13] https://www.cdc.gov/flu/about/season/flu-season-2017-2018.htm

[14] As part of the fitting we incorporate a time variable to further control for seasonality and any modest variations in exposed to risk between the individual calendar years

Group	Annuitant men	Annuitant women	Disabled retiree men	Disabled retiree women	Surviving beneficiary men	Surviving beneficiary women
Rating factors included	Age, ZIP+4, annuity amount, collar type	Age, ZIP+4, annuity amount, collar type	Age	Age	Age, ZIP+4, annuity amount	Age, ZIP+4, annuity amount
Life years of exposure	539,924	284,990	17,310	4,359	21,326	100,012
Number of deaths	23,597	9,685	769	132	1,413	5,979

Note: Greater data volumes are used when fitting the less granular models (i.e. where a subset of rating factors is used) – see our Data Underpinning ZIP+4 curves paper.

4.3 Our general function for mortality

For each group, we then fit a GLM of the following general form:

The predictors j are the longevity group (A to G as determined by ZIP+4), annuity amount and collar type

Main effect for each predictor: Additions depending on the value taken by each predictor j (can be negative)

Controls for mortality rate variations between calendar years, and is 0 for central year

$$logit(q_x | \overbrace{values\ of\ predictors}, j) = \underbrace{\textstyle\sum_i a_i x^i} + \overbrace{\textstyle\sum_j b_j} + \underbrace{\textstyle\sum_{i,j} c_{ij} x^{-i}} + \overbrace{\text{YOE}}$$

Main age function: A polynomial in age, x, with a small number of terms (typically 3 or 4) where i takes values in range $[-4, -3, \ldots, 3, 4]$

"Interaction" terms, whereby there is a small number of terms of the polynomial in age, x, which depend on the value taken by the predictor

For each group (annuitant men, etc.) we identify an "optimal" form by considering a range of possible polynomial functions of age (including the reciprocal of age, see "allowing for the compensation law" below) and potential interaction terms. For each possible model, the parameters are fitted by maximum likelihood estimation. The final model to use is chosen based upon the parsimony principle (i.e. minimizing the Akaike Information Criterion statistic across models subject to it while also meeting the goodness of fit requirements set out in section 5).

These forms are fitted to mortality rates on a *lives* basis. This is because the impact of affluence on longevity (i.e. that higher annuity amounts tend to have lighter mortality and are financially more significant at a plan level) is explicitly controlled for by the annuity amount variable within the functional form.

An example: Annuitant men, longevity group F, annuity of $15,000 p.a., white collar

The final model used for annuitant men where ZIP+4, annuity amount and collar type are available is set out below. For ease we have grouped terms to identify which part of the polynomial they influence.

$$logit(q_x|ZIP+4=j, Annuity=k, Collar=l) = a_0 + \frac{a_1}{x} + \frac{a_2 + c_{3l}}{x^2} + \frac{a_3}{x^3} + \frac{a_4 + c_{1j} + c_{2k}}{x^4}$$

With the coefficients for age being:

a_0	a_1	a_2	a_3	a_4
109.0	-2.966×10^4	3.063×10^6	-1.450×10^8	2.604×10^9

Note that the magnitude of these reflects the power of age term they are being divided by (for example at age 100 the $\frac{1}{x^4}$ term divides through 10^8).

In this case both the ZIP+4 longevity group and the annuity band moderate the quartic term in reciprocal of age, while collar type moderates the quadratic term in reciprocal of age. This means that the impacts of annuity amount and ZIP+4 attenuate faster with age than the impact of collar type. For our example man:

ZIP+4 longevity group F ($j = F$ so c_{1F})	Annuity amount $15,000p.a. (annuity band $k = 4$, c_{24})	White collar ($l =$ "W", so c_{3W})
-3.280×10^{-6}	-1.933×10^{-6}	-967.2

Hence, the mortality rates (within the age range 65-95) in our example are determined by:

$$logit(q_x) = 109.0 - \frac{29{,}660}{x} + \frac{3{,}062{,}000}{x^2} - \frac{145{,}000{,}000}{x^3} + \frac{2{,}596{,}000{,}000}{x^4}$$

NB: Throughout this example the coefficients have been rounded to 4 significant places. The full decimals are used to produce mortality rates

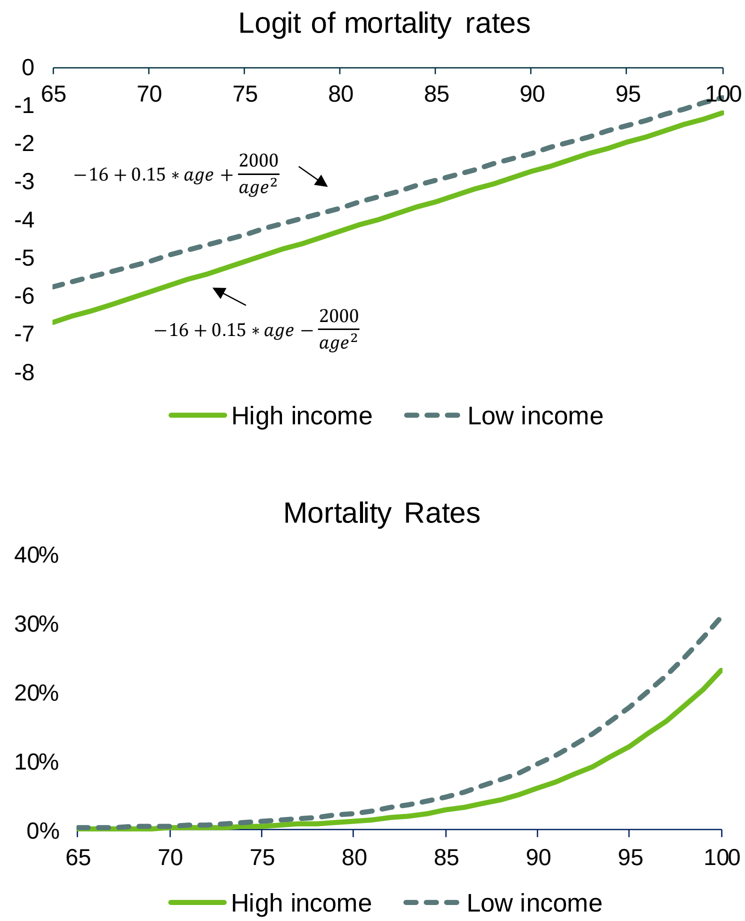
Allowing for the compensation law

The “compensation law” of mortality refers to how higher rates of mortality among certain populations (such as blue collar workers compared with white collar workers) are compensated for by a slower increase with age. Those who survive in the heaviest mortality group will tend to represent the healthiest subset of that group, creating a selection effect which reduces the *relative* differences in mortality rates between different populations with age. (There is also a natural limit imposed on the ratio of mortality rates which declines with age, since mortality rates must lie between 0 and 1.)

A key consequence of this is that applying a simple multiplier to off-the-shelf tables (like the RP-2014, Pub-2010 or Pri-2012) to match the experience of a particular plan, or a particular group within our dataset, is likely to misstate mortality at younger ages and older ages.

Within a GLM framework, the compensation law can be readily allowed for by including terms of how the impact of different rating factors varies with age. These terms (known technically as “interactions”) typically take the form of a multiple of a function of the reciprocal of age, with the multiple dependent on the value taken by the rating factor. A reciprocal of age is chosen as it naturally achieves convergence in relative mortality rates while permitting the absolute differences in mortality to increase with age, consistent with the compensation law. This is illustrated in Figure 6.

Figure 6: Achieving the compensation law of mortality



4.4 Including less granular curves

For some plan participants, not all data will be available – for example collar type may not be known, or ZIP+4 is not provided. Our general principle is that we want to use as much data as possible on an individual to provide as good an estimate of mortality for the individual as possible. Accordingly, we calibrate GLMs to different combinations of available information as illustrated by the table below. This means that there is always a curve available to use with the participant.

Each row in the table is a separate GLM for each of annuitant men/women, disabled retiree men/women and surviving beneficiary men/women of a deceased retiree, giving rise to 26 GLMs and a total of 306 VitaCurves. Each of these GLMs is subject to robust goodness of fit tests.

We believe that ZIP+4 can be readily sourced for the vast majority of pension plan participants. However, we accept that it may not be available for some plan participants and so have also created a clustering based upon ZIP code for use with these individuals. We have clustered the ZIP codes using the techniques described earlier using for each ZIP the modal socio-economics across all ZIP+4 codes within that ZIP. As many ZIP codes cover a larger area (and potentially many tens of thousands of lives), this is less informative of longevity and fewer distinct groups are possible; 6 for men and 5 for women. This leads to further 224 VitaCurves available within our model, for a total of 530 VitaCurves.

<table>
<tr><th colspan="3">Which variables included</th><th colspan="6">Number of resulting VitaCurves</th></tr>
<tr><th>ZIP+4</th><th>Annuity amount</th><th>Collar type</th><th>Annuitant man</th><th>Annuitant woman</th><th>Disabled retiree man</th><th>Disabled retiree woman</th><th>Surviving beneficiary (man)</th><th>Surviving beneficiary (woman)</th></tr>
<tr><td>✓</td><td>✓</td><td>✓</td><td>84</td><td>36</td><td colspan="2" rowspan="7">Not calibrated owing to sparseness of data</td><td colspan="2">n/a</td></tr>
<tr><td>✓</td><td>✓</td><td>×</td><td>42</td><td>18</td><td>14</td><td>18</td></tr>
<tr><td>✓</td><td>×</td><td>✓</td><td>14</td><td>12</td><td colspan="2" rowspan="2">n/a</td></tr>
<tr><td>×</td><td>✓</td><td>✓</td><td>12</td><td>6</td></tr>
<tr><td>✓</td><td>×</td><td>×</td><td>7</td><td>6</td><td>7</td><td>6</td></tr>
<tr><td>×</td><td>✓</td><td>×</td><td>6</td><td>3</td><td>2</td><td>3</td></tr>
<tr><td>×</td><td>×</td><td>✓</td><td>2</td><td>2</td><td colspan="2">n/a</td></tr>
<tr><td>×</td><td>×</td><td>×</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
</table>

5 Ensuring the model is a good fit to the data

Having identified and chosen the main variables to include in the model and fitted the model, we then need to examine whether the model provides a good fit to the mortality experience of the pension plan members observed in our dataset. There are a variety of summary test statistics and measures that are useful to assess the goodness of fit of a model. A full assessment cannot rely on one number, and instead we consider the results for multiple tests applied to each curve to identify that a model is a good fit to the underlying data. These same statistics are also used as a mechanism for comparing the merits where there may be multiple candidate models, for example for how we capture the shape of mortality with age, or how a variable changes mortality rates at different ages.

5.1 Statistical tests

When shortlisting candidate models we initially focus on statistical criteria of goodness of fit. Key tests used include:

- **Akaike Information Criterion (AIC):** This measures the goodness of fit via the (log-) likelihood (i.e. a measure of how likely the observed data is under the fitted model), with a penalty based upon the number of variables used (i.e. to avoid "over-fitting" to the data).
- **Bayesian Information Criterion (BIC):** This is defined similar to the Akaike Information Criterion but uses an alternative penalty which is more punitive to the number of parameters. The BIC and AIC statistic have competing merits and so we would tend to refer to both, especially in borderline decisions.

The AIC and BIC metrics are overall goodness of fit measures. However, they may miss subsets of poor fit in the model. To control for this, we also use:

- **Hosmer & Lemeshow test:** This is a statistical test based upon splitting the data into risk groups (deciles) and comparing the square of the standardized differences in actual vs expected deaths within each decile. If the model is a good fit this will be consistent with an observation from a χ^2-distribution[15].

[15] Full details of the theory behind this test can be found in Applied Logisitic Regression (Hosmer & Lemeshow, Wiley Series in Probability & Statistics)

5.2 Actuarial goodness of fit tests

A number of other tests have been developed specifically for use when graduating mortality tables – to ensure that the tables are a good fit to the underlying data across the age spectrum and for subsets of the data[16]. These tests tend to focus on differences observed between the actual deaths experienced at each age, and those that would be expected if experience was in line with the fitted rates. To control for the variability in mortality rates at each age these differences are standardized by the square root of the expected number of deaths so that they represent the number of standard deviations the observed mortality lies from the fitted rates at each age. The tests we apply (at the 95% confidence level) to these "residuals" for each VitaCurve include:

- **Actual vs Expected deaths:** A confidence interval is placed around the observed residuals, which should include 0 indicating that the experience is consistent with the fitted rates
- **Chi-squared test:** Are the residuals "random"? If they are, the sum of their squares should be from a chi-squared distribution
- **Signs test:** A test to check the residuals are not biased to positive or negative values (i.e. that the VitaCurve is not systematically under- or over-estimating mortality rates). Under this test, the number of positive residuals should be consistent with an observation from a binomial distribution with a 50% incidence rate.
- **Runs test:** Are there groups of ages where the residuals are consistently positive or negative? If so, there are parts of the age range where the VitaCurve is under- or over-estimating mortality rates. This test looks at the number of groups ("runs") of ages where the residuals have the same sign and checks this is consistent with what would be expected if the residuals are "random".
- **Serial correlations:** This checks that the residuals at age x and age $x - j$ ($j > 0$) are absent of correlations, consistent with any residuals being randomly distributed about the fitted curve. We typically focus on $j = 1$.
- **Kolmogorov-Smirnov test:** A test to ensure that there is not a series of small differences between the fitted and observed mortality rates which aggregate to an unacceptably large cumulative deviation across the ages fitted to.

5.3 Goodness of fit at the plan level

The tests described above ensure that the VitaCurves are a good fit to the dataset as a whole. However, they will be ultimately applied to value the liabilities of individual pension plans. It is important, therefore, that the curves are predictive of experience at the plan level. To assess this, we consider the ratio of actual deaths to those predicted by the VitaCurves for each plan. We do this on an amounts basis i.e. weighting each life/death by the annuity amount in payment to ensure that we capture the financial materiality of the larger annuity amounts. Figure 7 shows the resulting A/E for each of the 86 plans with detailed ZIP+4 information. Reassuringly, there is an even split between the number of plans for which this above and below 100%.

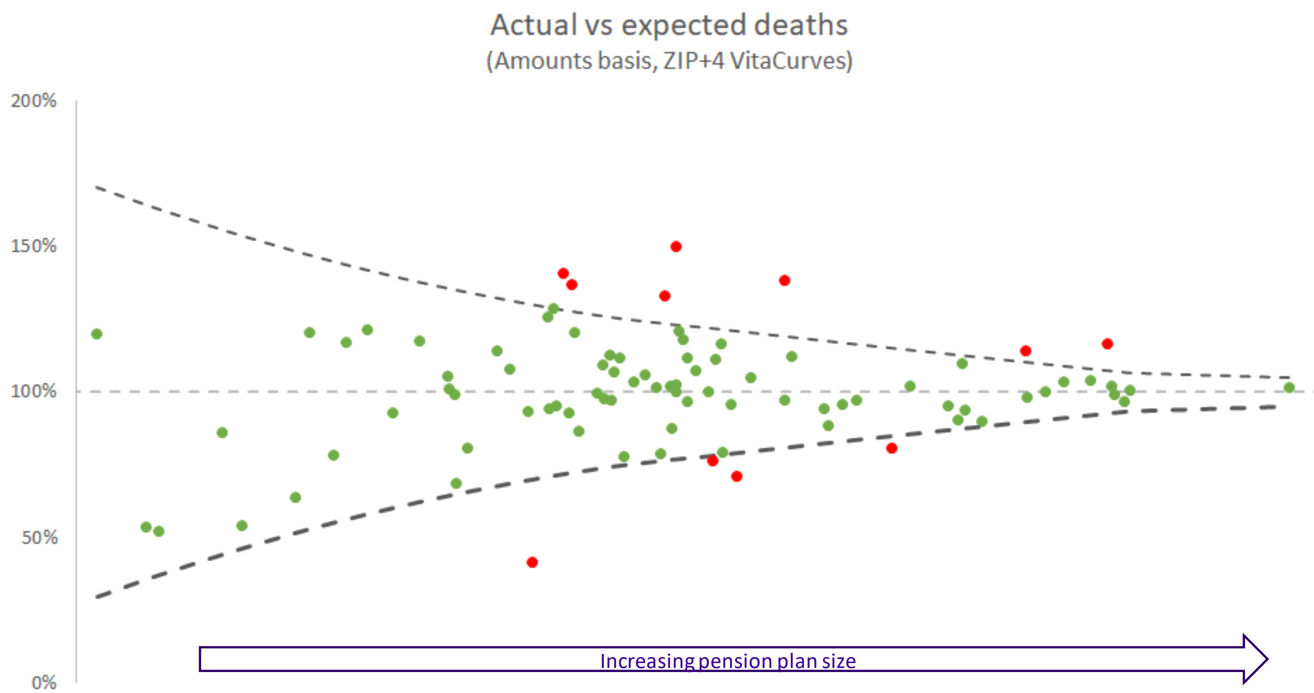
The dashed line in the figure represents a 95% confidence interval for the A/E ratios. The plans marked in green are those falling within the confidence interval (75) and those marked in red (11) fall outside. This distribution is markedly better than seen when we perform similar calculations for the recently-published Pri-2012 tables (using blue, white or all collar type tables based upon the dominant proportion in the plan[17]).

---

[16] See for example *A Practitioners Guide to Statistical Mortality Graduation* or *On graduation by mathematical formula* for detailed discussion of actuarial goodness of fit tests

[17] Using collar type tables where the proportion of participants of a blue or white collar exceeds 70%, per SOA guidance

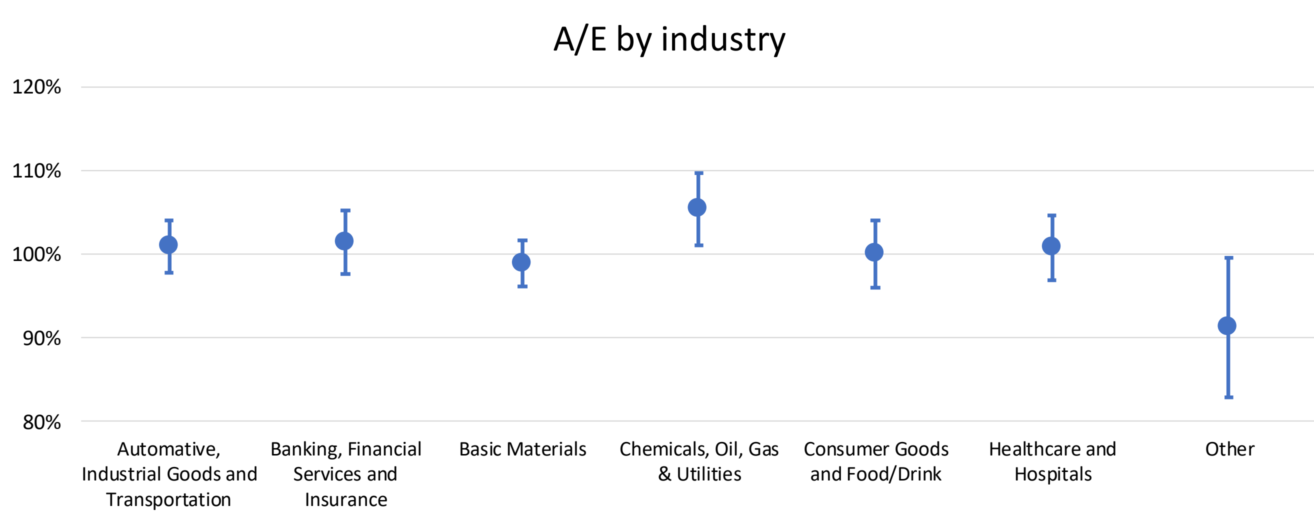
Figure 7:



5.4 Checking for residual effects

Our final check is to consider whether there are any systematic residual effects for variables we have not allowed for in our model. This helps to identify variables which may be useful to include in future editions of VitaCurves. For example, figure 8 considers the ratio of actual to expected deaths (on an amounts basis) for each industry in the dataset. Reassuringly, this is centered broadly on 100% for each industry suggesting that industry has little further to add to the model. This is consistent with our observations when considering which variables to include in the model.

Figure 8:



The potential exception to this appears to be Chemicals, Oil & Gas industry where the actual experience was heavier than suggested by VitaCurves. This is a diverse industry grouping including two large plans with reasonably heavy experience. Analysis from other territories indicates that this can sometimes be the case for certain companies in this sector. As we collect more data we will continue to monitor whether an industry factor in our model would be appropriate to capture the "outlier" industries. ("Other" in the above chart is a diverse group generally of smaller plans, with the overall experience being heavily influenced by one larger plan; we will also explore this group further as the dataset grows.)

6 Ensuring internal consistency

We also ensure that the resulting VitaCurves are internally consistent. For example, we would expect that:

- For a man and woman who are alike in all other ways, the male mortality should exceed the female mortality;
- As we increase a covariate value (e.g. affluence), but keep all other covariate values the same, the mortality should decrease (except possibly for the oldest old); and
- Former blue collar workers should have higher mortality rate than those who served as white collar workers (all other covariates alike).
- Disabled retirees should have higher mortality than annuitants.

While the results within any GLM should generally be internally consistent, the use of multiple GLMs to cover different combinations of available rating factors increases the risk that we may have some internal inconsistencies. We therefore assess the number of consistency issues in VitaCurves, including "unexpected" crossings (i.e. pairs of curves where the order of which has the highest mortality rates crosses at some point over the age spectrum) against specified criteria and tolerances. The criteria capture both the nature of consistency issue[18], and whether it is of concern[19]. With a large number of VitaCurves, it is likely that there will be a modest number of inconsistencies, and so tolerances are set on the number of inconsistencies identified. Breaching of certain thresholds is likely to lead to alternative GLMs being fitted. Where these thresholds are not breached, an experienced member of the team looks closely at the any inconsistencies identified before the VitaCurves are published.

7 Filling out the mortality table (extensions)

Our modeling provides mortality rates for the age range that we have fitted over. To value annuity benefits, we need to "fill out" the VitaCurves to cover the very oldest ages. We also fill out the VitaCurves down to younger ages so that they can be used with younger participants.

7.1 Completing the tables at the highest ages

Extending the VitaCurves to the highest ages requires a balance to be struck between a range of desirable features, including continuity, monotonicity, maintaining consistency between curves, biological reasonableness and smoothness.

[18] We require that VitaCurves should be monotonically increasing with age, similar granularity curves should be in the expected order (men above women) and less granular curves should be within the extremes of more granular curves

[19] Some inconsistencies may be fine on grounds of materiality (e.g. number of ages impacted, or in size of inconsistency)

The approach we have used builds upon a detailed study carried out by the High Age Mortality Working Party ("HAMWP") of the Institute and Faculty of Actuaries, published over a series of Working Papers[20]. For each VitaCurve the approach requires two steps:

1 Identify a natural "parent curve" to extend relative to

2 Extend the VitaCurve by exponentially decaying the gap between the VitaCurve and the parent "curve" at a chosen rate

To implement this approach, we need to identify a suitable parent curve and an appropriate rate of decay.

Choice of parent curves

For annuitant men/women and surviving beneficiary men/women, each VitaCurve converges to a parent of a VitaCurve for the same group which has been fitted to the whole dataset (i.e. allow for mortality rates to vary depending on an individual's age ("age only curve")).

For this to work, the "age only curves" have to have already been extended to older ages. Consistent with the proposals set out by the HAMWP, we do this by reference to a table fitted to the US national mortality data as sourced from the Center for Disease Control (CDC). We have created this table using the published CDC mortality rates (for 2015)[21] for ages up to 95 and then converging the mortality rate at 95 to a mortality rate of 63% at age 125 using non-linear interpolation[22,23].

The disabled retiree VitaCurve, which are fitted to younger ages, converge in to a "parent" of the annuitant curve of the same gender to inherit the shape of pension plan mortality at older ages.

Rate of decay

Based upon the rate of convergence of mortality rates at the oldest ages in the underlying data we have set the rate of decay to 10% (i.e. for every additional year of age, 10% of the remaining gap between the mortality rates is closed).

7.2 Providing mortality rates at the younger ages

We identify a population-level mortality rate at a young age (i.e. prior to the "accident hump"). The VitaCurves are then generally set to decrease linearly in the logit scale from the rate at the youngest age fitted to, to the population-level mortality at that younger age. This is a pragmatic approach for younger ages which reflects the broad linearity of the mortality rates within the logit scale and the low financial materiality of the choice of method.

The main exception to this approach is for the disabled retiree VitaCurves, where the curve is held constant at younger ages, consistent with the data we see at the younger ages.

---

[20] See CMI Working Papers 85, 100, 106 and 112. (Note there is also a dependency on Working Paper 78).

[21] CDC data available for men here and women here with associated documentation here

[22] $q_{125} = 63\%$ has been chosen consistent with studies that the force of mortality appears to flatten at around 1 at advanced ages; 125 chosen as the ultimate age to be consistent with the possibility of survival beyond 120 as believed to have been evidenced historically with the case of Mdme Jeanne Calment

[23] The non-linear interpolation uses a power-law interpolation to provide curvature to the mortality rates at the oldest ages and takes the form of $\mu_x = \left(\frac{125.5-x}{30}\right)^C \mu_{95.5} + \left(1-\left(\frac{125.5-x}{30}\right)^C\right)\mu_{125.5}$ for ages 95.5 to 125.5, where C is set to 1.25 which by inspection provides a reasonable progression of mortality. Underlying m_x values for CDC data at older ages used to determine $\mu_{95.5}$ using usual Uniform Distribution of Deaths ("UDD") assumption. Conversion of resultant μ_x to q_x required for VitaCurves performed using Boole's rule.

8 Want to know more?

If you have any questions on this technical document or would like to know additional details regarding our methods for fitting our US VitaCurves, please contact any of the team below. We would be delighted to hear from you.




Steven Baxter FIA
Head of Innovation & Development
Club Vita LLP
steven.baxter@clubvita.net

Dan Reddy FSA EA MAAA
CEO
Club Vita US
daniel.reddy@clubvita.net

Erik Pickett PhD FIA CERA
CPO
Club Vita US
erik.pickett@clubvita.net

James Maloney FIA
Head of Longevity Modelling
Club Vita LLP
james.maloney@clubvita.net

September 2019

For and on behalf of Club Vita LLP

Reliances and Limitations

In this paper (the “Research”), Club Vita LLP has provided an overview of the methodology used for the calibration of the first generation of US VitaCurves. The Research is based upon Club Vita LLP’s understanding of legislation and events as of August 2019 and therefore may be subject to change. Future actuarial measurements may differ significantly from the estimates presented in the Research due to experience differing from that anticipated by the demographic, economic or other assumptions. The Research should not be construed as advice and therefore not be considered a substitute for specific advice in relation to individual circumstances and should not be relied upon. Where the subject of the Research refers to legal; matters please note that Club Vita LLP is not qualified to give legal advice, therefore we recommend that you seek legal advice if you are wishing to address any legal matters discussed in this Research. Please be advised that Club Vita LLP (not its respective licensors) does not accept any duty, liability or responsibility regarding the use of the Research, except where we have agreed to do so in writing.

When read along with the “Zooming in on ZIP codes” and “Data Underpinning ZIP+4 VitaCurves", this paper complies with the relevant Actuarial Standards Board’s Actuarial Standards of Practice (ASOP) and Financial Reporting Council’s Technical Actuarial Standard (TAS) 100: Principles for Technical Actuarial Work.