Institute and Faculty of Actuaries

ISSN 2397-7213

Longevity Bulletin

From the Institute and Faculty of Actuaries

The machine learning issue

Issue 15

August 2023

Contents

Introduction by the Editor

The calculation of an actuarial value, whether in the context of pricing or reserving, generally requires the most accurate possible assumption regarding the contingent event in question: for most life and pensions actuaries that will be the event of death or, conversely, survival. The accuracy of the assumption depends in large part on correctly identifying and quantifying the mortality-related characteristics of the life in question.

Traditionally, age and sex have been the two key factors, with the concept of 'wealth implies health' allowed for implicitly via an appropriate amounts weighting of the analysis. But the heterogeneity remaining after allowing for those primary factors can be considerable, causing problems for pricing and reserving.

Driven largely by competitive pressures, along with the example already seen in retail non-life pricing which had been quick to spot the benefits of multi-factor analytical methods, life actuaries started to test the applicability of these models to mortality in the mid-2000s. The use of generalised linear models (GLMs) and survival models on annuity portfolios allowed retail annuity providers to introduce postcode into their pricing bases as a highly predictive proxy for socio-economic status. Insurers with richer data were able to use such methods to develop more predictive underwriting models using additional factors such as BMI, smoking status and medical history, if available.

At around the same time, the rise of the bulk purchase annuity market led to the requirement to be able to rate the mortality of small pension schemes lacking credible experience; this need was solved by the development of postcode-rating engines, calibrated from GLM analyses of large multi-scheme datasets.

By the mid-2010s, UK life insurers and pension consultancies, and some non-UK multinationals, took such progress for granted and started to think of the next steps. Machine learning offered a natural progression, given the machine learning successes seen in many other fields.

Over the last five or so years, machine learning has been tried in many life insurance contexts. The predictive boost of machine learning can be valuable, but this extra predictiveness is not 'free'. To be able to function well, machine learning generally requires very large data volumes. But the main disadvantage of machine learning is the reduction in transparency and communicability compared with GLMs or survival models.

For actuaries, the availability of a complex and very predictive modelling tool that is also difficult to interpret and to communicate is a great opportunity, strange though it may seem. Actuaries are able to understand what is going on 'inside', but we are also able to assess the output from a practical perspective: Do the results make sense? Can I use these results in my pricing or reserving models? What could go wrong? How can I best communicate the method, and its strengths and limitations?

Machine learning, or data science more generally, offers the actuarial profession the best of both worlds – a powerful suite of models and methodologies, along with the opportunity to 'add value' to the work of non-actuarial data scientists through many of these practical points.

We hope you find this machine learning issue of the Longevity Bulletin interesting, and that it illustrates how actuaries can use these, still relatively novel, techniques for the ultimate benefit of our insurance and pension fund stakeholders.

Matthew Edwards
Editor

Foreword by the President of the IFoA

While it's a truism that we operate in a constantly changing world, with each year bringing fresh challenges, many of these changes can actually be positive.

Actuaries, and the actuarial profession, seek to overcome fresh obstacles and new problems using the most recent 'tools' – whether those tools be methodologies, computer resources, new datasets, or anything else.

The rise of data science is a great example of how those three aspects have grown in parallel to some extent in recent years. That 'parallel' growth is not of course a coincidence, as each has fed off, and indeed fed, the others: more powerful computers and richer data allow us to try potentially more powerful methods.

The analytic opportunities of data science in general, and machine learning in particular, have not been wasted on actuaries. As we see in the articles in this issue, even within just the confines of mortality and longevity there are plenty of ways we can deepen our understanding of mortality risk through the deployment of these techniques.

The Institute and Faculty of Actuaries has moved accordingly, and since 2020 our Certificate in Data Science has provided IFoA members with an introduction to data science. In the three years the course has run so far, over 500 members have successfully passed and been awarded the Certificate. In addition, aspects of data science have been added to the IFoA's Fellowship exams, such as the introduction of practical papers in R for our Core Statistics (CS) subjects, and a module on machine learning in CS2. Some data-science related content has also been added to our later practice-area-specific specialist exams.

The IFoA intends to add further data-science related content to its Fellowship examinations in the future, including a specific data science route to Fellowship, with a greater emphasis on the practical application of data science skills. This is a powerful example of actuaries adapting so as to take advantage of new fields and techniques. With data science, we are not seeking to replace data scientists, but to ensure we can fully understand the various nuances of the subject in order to add value.

We can apply our understanding of aspects such as model risk and trend projection, or practical points such as an appreciation of how best to use the results and communicate the method, while also allowing data scientists the space to flourish.

Although ChatGPT is the main topic of conversation in tech and innovation forums this year, there is still some way to go before actuaries can claim to have exhausted the possibilities of machine learning. No doubt ChatGPT and equivalent AI systems can advise us on what stones we have left unturned!

My thanks to the authors and editors of these articles for this valuable edition of the Longevity Bulletin. In an ever-changing world, it's great to see a clear statement of some of the positive changes!

Matt Saker
President of the Institute and Faculty of Actuaries

Multivariable mortality modelling, survival analysis, and machine learning

John Ng, Director, Longevity Analytics at RGA

The modelling and management of mortality and longevity risks are essential for insurers, reinsurers, pension funds, banks and government agencies. More advanced models can yield a more accurate mortality estimation that can assist pricing, reserving, underwriting, solvency and profitability. In addition, a rich and robust modelling methodology provides a more comprehensive understanding of mortality risks. This is important for fostering design innovations for annuities, equity release, and protection products, as well as supporting the role insurers and pension funds play in financing longevity risks.

In this article I set out an overview of mortality modelling using survival analysis techniques, a discussion of machine learning methods, extensions to the generalised linear models (GLM), and some examples of applying these models. The focus will be on the use of individual-level data in models that incorporate a multitude of risk factors, rather than group-level or population-level data.

1. Mortality modelling and survival analysis

Traditionally, actuaries and demographers made extensive use of mortality rates, ie the probability of a group of similar lives dying in one year. In more recent times, mortality modelling has advanced considerably, with model development employing a statistical framework capable of statistical tests and confidence intervals. At the same time, there has been a significant increase in the volume, veracity, velocity and variety of data available for analysis, which encompasses policyholder data, demography, postcodes, electronic health records, lifestyle and credit scores.

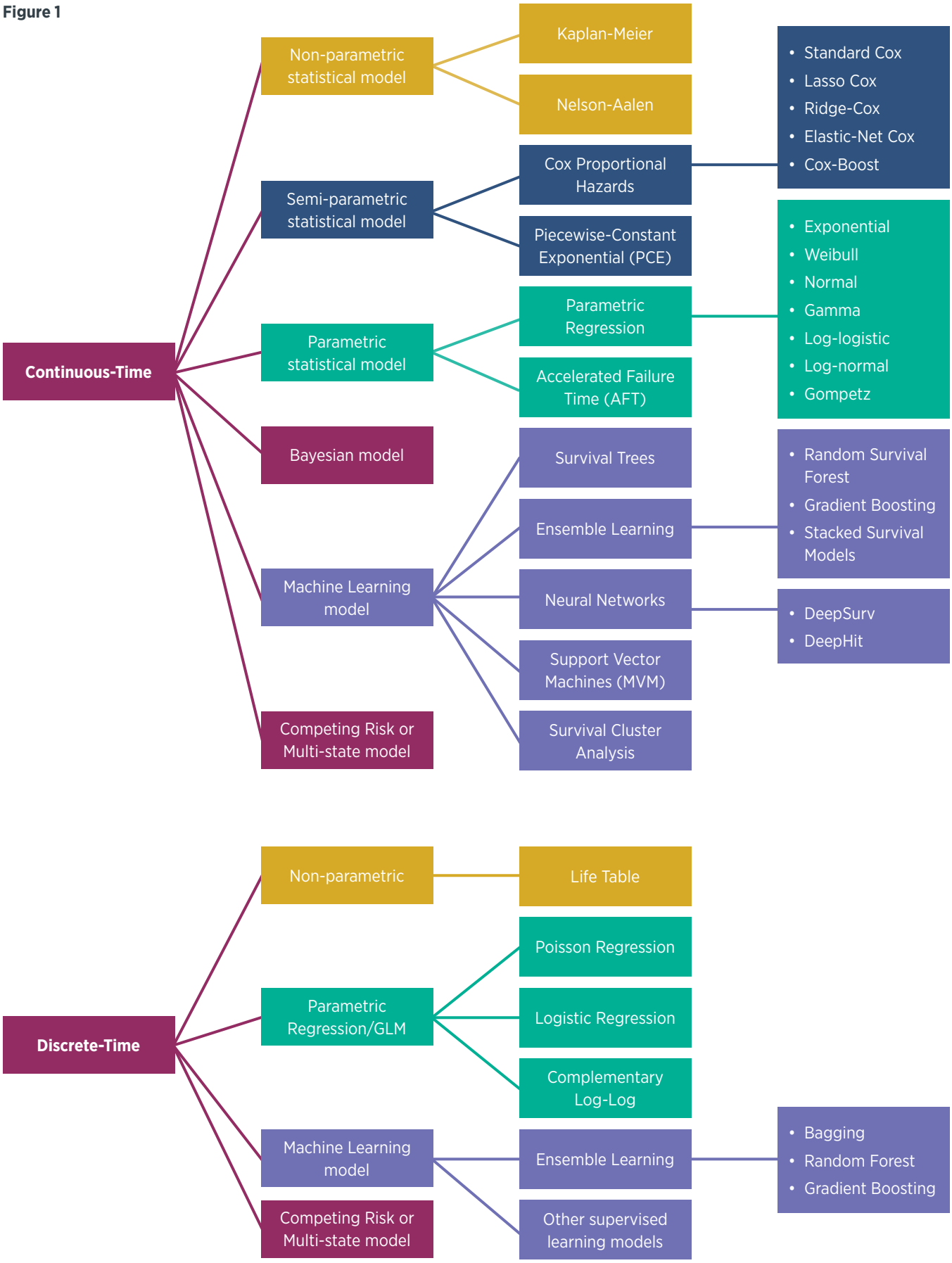
As insurers and pension schemes collect experience data on individual policyholders, this is essentially a form of longitudinal study because the individuals are observed for a period of time until a particular event of interest occurs (known as time-to-event or survival time). The main challenge of time-to-event data is the presence of incomplete observations. This occurs due to censoring, ie when a death event happens outside of the observation period. For this reason, many classical statistical models and machine learning methods could not transpose directly to time-to-event data.

Modelling time-to-event data requires a specific approach called survival analysis. It is used to predict the target variable of survival time until mortality, while accounting for censoring, and the presence of explanatory variables that may affect survival time (Rodríguez, 2007).

Survival analysis is a very useful tool in evaluating risks such as mortality, longevity, morbidity, lapses, and demographic factors (eg marriage, migration and fertility). As the name indicates, survival analysis has origins in the field of medical research to estimate the survival rate of patients after a medical treatment. It is also known as reliability analysis in engineering, duration analysis in economics, and event history analysis in sociology (Abbas et al., 2019).

Figure 1 shows a taxonomy of survival analysis models, which provides a holistic view of statistical and machine learning methods, categorised by continuous-time and discrete-time approaches (with separate charts for each).

Figure 1


Continuous-Time
Non-parametric statistical model
Kaplan-Meier
Nelson-Aalen
Semi-parametric statistical model
Cox Proportional Hazards
Piecewise-Constant Exponential (PCE)
Standard Cox
Lasso Cox
Ridge-Cox
Elastic-Net Cox
Cox-Boost
Parametric statistical model
Parametric Regression
Accelerated Failure Time (AFT)
Exponential
Weibull
Normal
Gamma
Log-logistic
Log-normal
Gompetz
Bayesian model
Machine Learning model
Survival Trees
Ensemble Learning
Neural Networks
Support Vector Machines (MVM)
Survival Cluster Analysis
Random Survival Forest
Gradient Boosting
Stacked Survival Models
DeepSurv
DeepHit
Competing Risk or Multi-state model
Discrete-Time
Non-parametric
Life Table
Parametric Regression/GLM
Poisson Regression
Logistic Regression
Complementary Log-Log
Machine Learning model
Ensemble Learning
Other supervised learning models
Bagging
Random Forest
Gradient Boosting
Competing Risk or Multi-state model

1.1 Continuous-time survival analysis

Survival analysis theory focuses on two key concepts in continuous time:

a. the survival function *S(t)*, ie the probability of being alive just before duration *t*

b. the hazard function *h(t)*, ie the instantaneous death rate at time *t*, also known as the force of mortality by actuaries

There is a one-to-one relationship between the hazard function and the survival function. Whatever functional form is chosen for the hazard function, one could use it to derive the survival function. The integral of the survival function then gives the expectation of life, ie mean of survival time (Rodríguez, 2007).

1.1.1 Non-parametric models

The non-parametric methods are simple and require no assumptions on distributions. The Kaplan-Meier estimator, also known as the product limit estimator, provides an empirical estimate of the survival function. The Nelson-Aalen estimator approximates the cumulative hazard function. As the sample size gets very large, these two estimators are asymptotically equivalent (Jenkins, 2005). Kaplan-Meier and Nelson-Aalen are univariable methods and likely to be less predictive; therefore, considering multivariable methods is recommended if multiple explanatory variables are available.

1.1.2 Semi-parametric models

In the semi-parametric category, the Cox proportional hazard model was proposed by Cox (1972) in perhaps the most often cited article on su rvival analysis. The hallmark of the Cox model is that it allows one to estimate the relationship between the hazard function and explanatory variables, without having to make any assumption on the baseline hazard function. Proportional hazards modelling assumes that the ratio of the hazards for any two individuals is constant over time. The fact that the hazards are proportional is helpful in making interpretations, such as when identifying the better treatment in medical trials or analysing loadings of risk factors in underwriting. The Cox model can also be generalised to handle time-varying covariates and time-dependent effects (Rodríguez, 2007).

The Piecewise-Constant Exponential (PCE) model is another example of semi-parametric continuous-time model and can be seen as a special type of proportional hazards model (Jenkins, 2005). When the time axis is partitioned into a number of intervals in a PCE model, it assumes that the baseline hazard is constant within each interval. The advantage is that one does not have to impose the overall shape of the hazard function in advance. Another useful property of the PCE model is its equivalence to a certain Poisson GLM model; this will be discussed later.

1.1.3 Parametric models

Parametric statistical methods assume that survival time follows a particular theoretical distribution (Wang et al., 2019). Commonly used distributions include exponential, Weibull, Normal, Gamma, log-logistic, log-normal and Gompertz (Jenkins, 2005). If the survival time follows the assumed distribution, resulting outcomes are accurate, efficient and easy to interpret, but if the assumption is violated parametric models can give sub-optimal results.

Another approach in the parametric category is the accelerated failure time (AFT) model. AFT assumes a linear relationship between the log of survival time and the explanatory variables. The effect of variables is to accelerate or decelerate the life course. The Weibull model is the only model that satisfies both proportional hazards and AFT assumptions (Rodríguez, 2007).

1.1.4 Other continuous-time models

As discussed in *Bayesian Survival Analysis* (Ibrahim et al., 2001), Bayesian approaches can be applied to survival models, including parametric, proportional and non-proportional models. Interpretability is a strength of Bayesian modelling.

Other examples of survival models include machine learning methods, which will be discussed in **Section 2**, as well as competing-risk and multi-state models (Jenkins, 2005).

1.2 Discrete-time survival analysis

The survival analysis techniques discussed in the previous section assume continuous measurement of time. Although it is natural to consider time as a continuous variable, in practice observations are often on a discrete time scale, such as days, months or years (Jenkins, 2007). An advantage of discrete-time modelling is the embedding of the GLM framework.

Interestingly, the PCE model is equivalent to a GLM Poisson log-linear model for discretised pseudo-data, when the death indicator is the response and the log of exposure times is the offset (Rodríguez, 2007). The likelihood function of PCE and independent Poisson observations happen to coincide and would therefore lead to the same estimates.

Generally, the choice of GLM for survival analysis depends on the nature of data (Rodríguez, 2007):

i. If data is continuous and if one is willing to assume hazard is constant in each interval, the Poisson GLM is appropriate as it allows use of partial exposures

ii. If data is truly discrete, logistic regression is recommended

iii. If data is continuous but only observed in grouped form, the complementary log-log link is preferable.

2. Machine learning models for survival analysis

In recent years, machine learning models have achieved success in many areas. This is due to built-in strengths that include higher prediction accuracy, ability to model non-linear relationships, and less dependence on distribution assumptions. Nevertheless, some machine learning algorithms bring notable weaknesses as well, such as difficulty in interpretation, sensitivity to hyperparameters, and a tendency to overfit. Dealing with censored data presents the biggest challenge for using machine learning in survival analysis. A deeper understanding of survival analysis, and how machine learning can overcome the challenge of censored data, is required in order to effectively adapt it to mortality modelling.

The discussion below starts with regularisation – a versatile machine learning technique applicable to many approaches, including GLM and classical survival models. The total range of machine learning models is vast; therefore, I look just at continuous-time models and deliberately exclude some notable discrete-time approaches, for instance support vector machines and random forest. Discrete-time supervised machine learning models are discussed in *Modelling Discrete Time to Event Data* (Tutz and Schmid, 2018), while extensions of GLM are discussed in **Section 3**.

2.1 Regularisation

Regularisation is a technique used to simplify a model and reduce overfitting by adding penalties or constraints to the model-fitting problem. The three main types of regularisation are:

i. Ridge, also known as Tikhonov or L2 regularisation, adds a penalty term based on the squared value of coefficients. It reduces the size of coefficients and deals with correlations between features simultaneously.

ii. Lasso (least absolute shrinkage and selection operator), also known as L1 regularisation, adds a penalty term based on the absolute value of coefficients. In contrast to Ridge, Lasso can shrink coefficients to zero, which means it can perform automatic variable selection. Extensions of Lasso include Group Lasso, Fused Lasso, Adaptive Lasso and Prior Lasso.

iii. Elastic net linearly combines the Ridge and Lasso penalty terms.

2.2 Cox models

Introducing regularisation into Cox proportional hazard models provides us with a form of machine learning – the resulting models include Ridge-Cox (Verweji and Van Houwelingen, 1994), Lasso-Cox (Tibshirani, 1997), and Elastic Net-Cox (Simon et al., 2011).

The Cox-Boost method (Binder and Schumacher, 2008) incorporates gradient boosting machines in Cox models. It is useful on high-dimensional data and considers some mandatory variables explicitly in the model.

2.3 Survival tree

Survival trees are classification and regression trees (CART) specifically designed to handle censored data (Gordon and Olshen, 1985). The data is recursively partitioned based on a splitting criterion and objects with similar survival times are grouped together. This approach is easier to interpret and does not rely on distribution assumptions.

2.4 Random survival forest and other ensemble methods

In machine learning, ensemble learning is a method that takes a weighted vote from multiple models to obtain better predictive performance than could be obtained from any of the constituent models alone. Common types of ensembles include bagging, boosting and stacking.

Bagging survival trees involves taking a number of bootstrap samples from the survival data, building a survival tree for each sample, and then averaging the tree nodes' predictions (Hothorn et al., 2004).

Random survival forest is similar to bagging, but random forest uses only a random subset of the features for selection at each tree node. This helps reduce the correlation between trees and improves predictions. Random survival forest does not depend on distribution assumptions and can be used to avoid the proportional hazards constraint of a Cox model (Ishwaran et al., 2008).

Boosting combines a set of simple models into a weighted sum and is iteratively fitted to the residuals based on the gradient descent algorithm. Hothorn et al. (2006) proposed gradient boosting to account for censored data.

Stacking combines the output of multiple survival models and runs it through another model. Wey et al. (2015) proposed a framework of stacked survival models that combines parametric, semi-parametric and non-parametric survival models. This approach has performed well by adaptively balancing the strengths and weaknesses of individual survival models.

2.5 Artificial neural networks

Artificial neural networks (ANN) consist of layers of neurons interconnected as a network to solve optimisation problems. The adjective 'deep' in deep learning refers to the use of multiple layers in the network. Neural networks and survival forests are examples of non-linear survival methods.

The initial adaptation of survival analysis to neural networks sought to generalise Cox with only one single hidden layer (Farragi and Simon, 1995). Katzman et al. (2018) later proposed DeepSurv, a deep feed-forward neural network generalising the Cox proportional hazards model. It has the advantage of not requiring a priori selection of covariates, by learning them adaptively.

DeepHit is a deep neural network that learns the distribution of survival times directly (Lee et al., 2018). Unlike parametric approaches, it makes no assumption of the underlying stochastic processes and allows for the relationship between covariates and risk to change over time. DeepHit can be used for survival datasets with a single mortality risk as well as multiple competing risks.

3. Extensions and enhancements of GLM

GLM is a popular tool in survival analysis due to its versatility, interpretability, predictive power and availability in many software packages. Section 1 demonstrated that GLM Poisson, Logistic and C-Log-Log models can perform survival analysis. However, GLMs elicit two common negative views: they are restricted by distribution assumption, and they do not account for non-linear relationships, which reduces predictive performance.

The first view is refutable because, as discussed in **Section 1**, a GLM is merely a device to derive the underlying survival model, so the model is not restricted by distribution assumptions of GLM.

The second issue can be mitigated using approaches such as these to extend or enhance GLM. Note that these are not restricted to survival analysis and can be applied to GLMs in general. Some practitioners would view these as ways to combine the advantages of GLMs (for instance, interpretability) with the power of machine learning:

1. **Generalised additive model (GAM)**: GAM is a GLM in which one or more of the predictors depends linearly on some smooth functions, which is useful to capture non-linear patterns. Examples of smooth functions are cubic splines and fractional polynomials. This approach allows much more flexible models.
2. **Generalised linear mixed model (GLMM)** – The GLMM extends the GLM by incorporating random effect terms. GLMMs are also referred to as frailty models (Tutz and Schmid, 2018).
3. **Regularisation** such as elastic net to handle multicollinearity and reduce overfitting.
4. **Automatic variable selectio**n using Lasso or elastic net. This can help identify influential risk factors efficiently rather than using stepwise selection, especially when the number of possible predictors is large.
5. **Identification of predictive interaction terms** with the help of machine learning, such as decision trees or random forest. If interpretability is important, it is preferable to keep the interaction terms relatively simple, rather than incorporating an influential yet hard-to-interpret 'blackbox' sub-model, such as a neural network, into a GLM.
6. **Dimension reduction**, by using unsupervised machine learning techniques, if there is a very large number of variables relative to the number of observations.

4. Applications in mortality modelling

Tedesco et al. (2021) constructed machine learning models to predict all-cause mortality in a two- to seven-year time frame in a cohort of healthy older adults. The models were built on features including anthropometric variables, physical and lab examinations, questionnaires and lifestyle factors, as well as wearable data. Random forest showed the best performance, followed by logistic regression, AdaBoost and decision tree. Additional insights could be extracted to gain understanding on healthy ageing and long-term care.

Using the MIMIC-III dataset on long-term mortality after cardiac surgery and the AUC metric, the researchers observed the order of model performance, from highest to lowest, to be AdaBoost, logistic regression, neural network, random forest, Naïve Bayes, XGBoost, bagged trees and gradient-boosting machine (Yu et al., 2022).

The OpenSAFELY paper (Williamson, 2020) applied the multivariable Cox model to analyse data from 17 million patients in England and subsequently identified a range of risk factors for Covid-19 mortality This was instrumental in helping to identify high-risk population subgroups, as Dan Ryan describes elsewhere in this Bulletin. Later that year, RGA (Ng et al., 2020) published a paper that cross-compared an all-cause mortality model with OpenSAFELY's Covid-19 model in a parallel and multivariable way. This revealed insights on excess mortality risk from certain factors, which were useful to actuaries and underwriters. Six months later, the OpenSAFELY team published another paper (Bhaskaran et al., 2021) analysing Covid-19 and non-Covid-19 mortality odds ratios, by using logistic regression. The team produced results that were very consistent with RGA's.

Conclusion

The goal of mortality modelling is to predict and understand mortality and longevity. This article provides a survey and taxonomy of mortality modelling under the survival analysis framework, structured by continuous-time and discrete-time, as well as statistical methods and machine learning. The choice of model depends on the nature of the data and the purpose – whether it is solely about predictive accuracy or if interpretability is important.

Due to the increasing availability of data, technology and development in survival analysis and machine learning, financial services providers, such as insurers and pension funds, can leverage advances in these areas to provide financial protection more effectively to more people.

References

Abbas, S.A., Subramanian, S., Ravi, R., et al. (2019). *An introduction to survival analytics, types, and its applications.* **https://www.intechopen.com/chapters/64244** [Accessed 2 Feb 2023.]

Bhaskaran, K., Bacon, S., Evans, S.J.W., et al. (2021). Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. *The Lancet Regional Health - Europe*, 6: 100109. **https://doi.org/10.1016/j.lanepe.2021.100109**

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9. **https://doi.org/10.1186/1471-2105-9-14**

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2): 187-220. **http://www.jstor.org/stable/2985181**

Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine*, 14(1): 73-82. **https://doi.org/10.1002/sim.4780140108**

Gordon, L. and Olshen, R.A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports*, 69(10): 1065-9.

Hothorn, T., Lausen, B., Benner, A., et al. (2004). Bagging survival trees. *Statistics in Medicine*, 23(1): 77-91. **https://doi.org/10.1002/sim.1593**

Hothorn, T., Buhlmann, P., Dudoit, S., et al. (2006). Survival ensembles. *Biostatistics*, 7(3): 355-373. **https://doi.org/10.1093/biostatistics/kxj011**

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., et al. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3): 841-860. **https://doi.org/10.1214/08-AOAS169**

Jenkins, S.P. (2005). *Survival analysis.* **https://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf** [Accessed 2 Feb 2023.]

Katzman, J.L., Shaham, U., Cloninger, A., et al. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18. **https://doi.org/10.1186/s12874-018-0482-1**

Lee, C., Zame, W.R., Yoon, J. et al. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). **https://doi.org/10.1609/aaai.v32i1.11842**

Ng, J., Bakrania, K., Falkous, C., et al. (2020). *COVID-19 mortality by age, gender, ethnicity, obesity, and other risk factors: a comparison against all-cause mortality*. RGA, 18 December. **https://www.rgare.com/knowledge-center/media/research/covid-19-mortality-by-age-gender-ethnicity-obesity-and-other-risk-factors** [Accessed 2 Feb 2023.]

Rodríguez, G. (2007). Lecture notes on generalized linear models. **https://grodri.github.io/glms/notes/** [Accessed 18 July 2023.]

Simon, N., Friedman, J., Hastie, T., et al. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5): 1-13. **https://doi.org/10.18637/jss.v039.i05**

Tedesco, S., Andrulli, M., Larsson M.A., et al. (2021). Comparison of machine learning techniques for mortality prediction in a prospective cohort of older adults. *International Journal of Environmental Research and Public Health*, 18(23): 12806. **https://doi.org/10.3390/ijerph182312806**

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4): 385-395. **https://doi.org/10.1002/(sici)1097-0258(19970228)16:4%3C385::aid-sim380%3E3.0.co;2-3**

Tutz, G. and Schmid, M. (2016). *Modeling discrete time-to-event data*. New York: Springer.

Verweji, P.L. and Van Houwelingen, H.V. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24): 2427-36. **https://doi.org/10.1002/sim.4780132307**

Wang, P., Li, Y. and Reddy, C.K. (2019). Machine learning for survival analysis: a survey. *ACM Computing Surveys*, 51(6). **https://doi.org/10.1145/3214306**

Wey, A., Connett, J. and Rudser, K. (2015). Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3): 537-49. **https://doi.org/10.1093/biostatistics/kxv001**

Williamson, E.J., Walker, A.J., Bhaskaran, K., et al. (2020). Factors associated with COVID-19-related-death using OpenSAFELY. *Nature*, 584: 430-6. **https://doi.org/10.1038/s41586-020-2521-4**

Yu, Y., Peng, C., Zhang, Z., et al. (2022). Machine learning methods for predicting long-term mortality in patients after cardiac surgery. *Frontiers in Cardiovascular Medicine*, 9. **https://doi.org/10.3389/fcvm.2022.831390**

John Ng

John Ng is Director, Longevity Analytics at RGA, where he is responsible for longevity data analytics, predictive modelling and tools for pricing and risk management, as well as assumption initiatives. He has worked in various areas of the (re)insurance market, including pension risk transfer, life and health research, and general insurance pricing. Previously he practised as a retail and hospital pharmacist. John is the current Chair of the IFoA Health and Care Research Sub-Committee, Deputy Chair of the IFoA Data Science community and a representative on the International Actuarial Association's Data Analytics Forum.

Practical aspects of data science

Valerie du Preez, Patrick Moehrke and Lara van Heerden, Actuartech

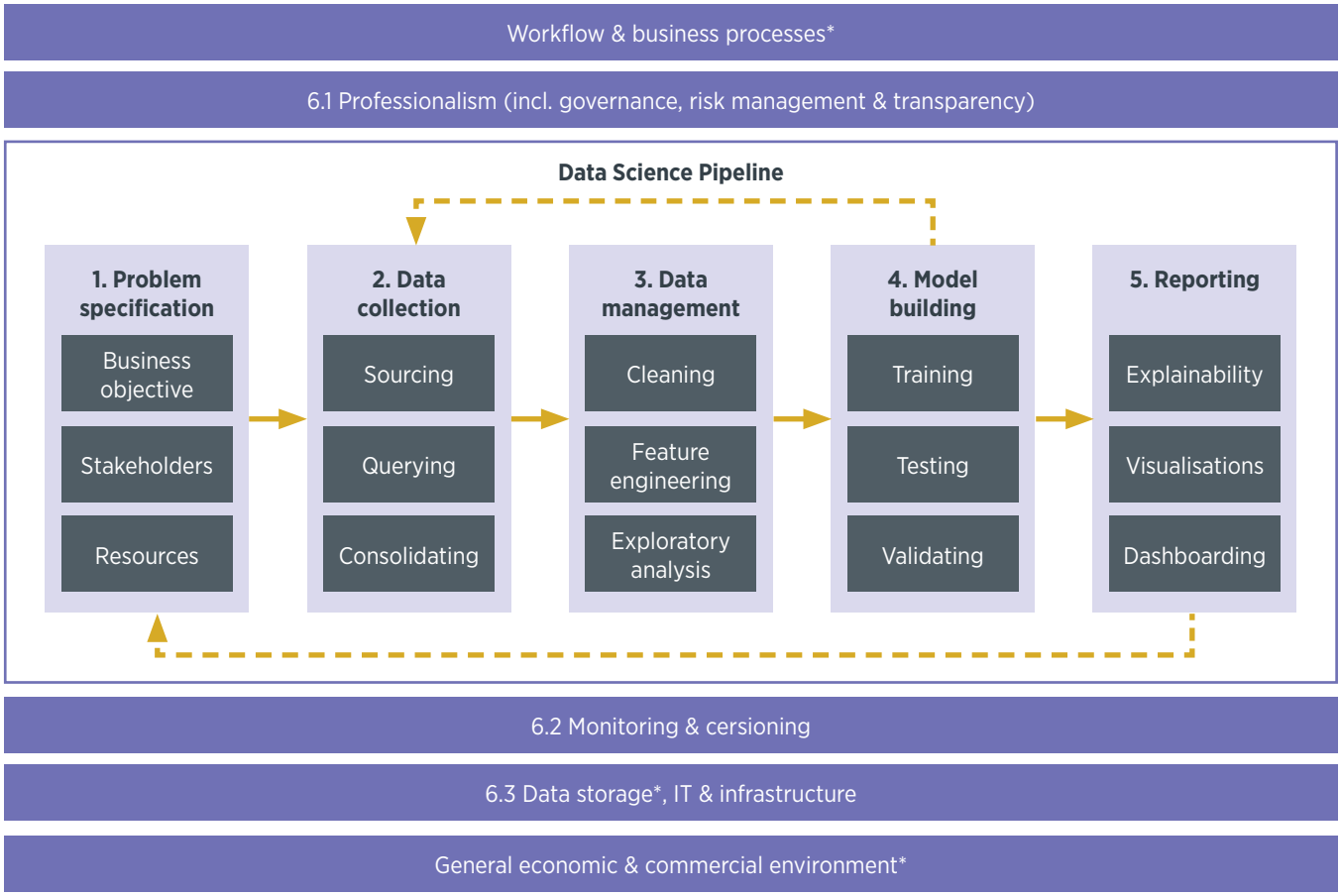
Introduction

Data science and modern analytical techniques are increasingly being utilised by the actuarial community, according to recent benchmarking by Actuartech on the adoption and use of data science (Actuartech, Reacfin and Synpulse, 2021) and modern infrastructure (Actuartech, 2023) within actuarial teams.

Embedding these techniques within existing actuarial processes can be particularly effective through new ways of visualising data, more sophisticated modelling approaches, and process flows.

In this article, we explore practical considerations and recommend actions to take in respect of different phases of an actuarial solution which, together with the actuarial control cycle, includes the data science pipeline and the supporting infrastructure and process components. *Figure 1* below shows an example structure to follow, and this article touches on most of these components. We also present an exploratory mortality experience analysis we performed to illustrate how certain considerations and actions look in practice.

Figure 1: *Actuartech Actuarial Solution Framework, which contains the actuarial control cycle (Bellis et al., 2010), the data science pipeline, and wider considerations. Note that we will not be explicitly discussing elements marked with an asterisk.*



1. Problem specification

As well as considering the specific objectives for analysing mortality experience in the context of available data, the actuarial function also typically aims to balance cost, complexity and accuracy of the solution. In relation to actuarial solution development, we are seeing actuarial teams incorporating solution design and development principles, strategic infrastructure decisions, policies and resource considerations (people, skills and infrastructure) in the problem specification phase.

In the investigation we present,[1] we performed an experience analysis investigation by comparing actual mortality against expected mortality for the annuity book of a life insurer. The expected mortality rates are taken from the PA(90)f (IFoA, 1990a) and PA(90)m (IFoA, 1990b) tables. As part of the investigation, we used an A/E model to identify to what extent the expected base mortality would need to change to better predict mortality for the book of lives.

Furthermore, we wanted to develop a predictive model to support pricing or valuation assumption setting, and the analysis we performed opened further business-related questions (not discussed here), including:

- To what extent did the actual experience deviate from expected?
- Is there evidence to suggest the underwriting process requires changing?
- Are the pricing and valuation bases still appropriate?
- Which areas of our modelling process and infrastructure requires updating?

When embarking on an investigation piece using new techniques, one must be mindful of the resources, including availability of the team, skillsets and feasibility given current deliverables. By expanding the aspects of the investigation to other domain experts (eg data scientists, data engineers and IT specialists), appropriate review and validation can be embedded in the process.

2. Data collection

During the data collection phase, it is important to consider what relevant data is available, internally or externally, whether it has suitable features, the quality, how much data is available, and whether or not the data can be enhanced. You must also ensure that legal and ethical considerations, including data protection requirements, are addressed.

For the example experience analysis, the following data inputs were sourced and queried using SQL, Python and Excel:

- Consolidated actual experience data, spanning three years at a per policy level
- Assumptions for calculating the expected level of deaths, including life tables
- Other important factors such as valuation date and information about important historical influences and factors, for example relevant information about Covid cycles.

It would also be helpful to consider additional data (where ethically available) to enhance existing data sources eg data from reinsurers and population statistics. This is particularly useful where data is not sufficient for a credible analysis.

3. Data management

Data management is an iterative process of transforming raw data into more informative and suitable variables. Tasks and techniques could include the following:

- **Cleaning** the data, for example by removing duplicates and handling missing values
- Creating new variables that might be predictive based on the dataset through **feature engineering**
- Reducing the number of explanatory variables through **dimensionality reduction**, for example by principal component analyses
- **Clustering** to create groups of similar data points
- Performing **exploratory data analyses** to better understand the data, for example through visualisations.

It could also be useful to perform preliminary model building where, for example, data is passed through a basic modelling pipeline, which comprises manipulating the data into the required form and fitting a generalised linear mode (GLM) (Dobson and Barnett, 2008) to test suitability.

Some of these tasks are discussed further below.

3.1 Data cleaning

To ensure a robust analysis, the quality of the data needs to be assessed to determine whether it is fit-for-purpose and the extent to which data cleaning is required. Data-cleaning techniques include:

- Identifying and handling missing, 'Null', 'not available' or incomplete values
- Performing reasonability checks (eg is the death date later than the inception date and is the age sensible?)

1 | Please note that this is a simplified indicative example, exploring certain practical considerations when incorporating data science in experience analysis, and the packages, findings, data or results presented should not be used to make any decisions.

- Investigating whether there are any duplicates or omissions in the data, and removing duplicates where appropriate
- Comparing data with that from previous investigations and independent records
- Visualising key summary statistics.

The aim of such checks is to identify errors and inconsistencies effectively and remove these to achieve a clean set of data that is appropriate for the investigation. Additionally, consider what specific tools and techniques will be most effective for achieving this aim, and ensure compliance with professional standards, such as TAS100 (FRC, 2023) in the UK, and data protection regulation.

In the example experience analysis, we applied the techniques above using Python. In our case, it was feasible to handle missing and nonsensical values on a case-by-case basis. We identified whether information could be retrieved elsewhere, or if the data point had to be removed altogether, and acted accordingly.

As measuring the cleanliness of data is often difficult, model accuracy and stability could be used to inform, for example, further data cleaning work required, making this a potentially iterative process step.

3.2 Feature engineering and exploratory analysis

The aim is to focus on those variables that are most relevant to the analysis and of sufficient data quality to support any conclusions drawn. In some cases, it is appropriate and necessary to derive or engineer 'synthetic' variables based on the raw data. Practical aspects inform choices made at this stage, including the compatibility of the data, the types of model being used for the analysis, and the suitability of the variables chosen for automation. As best practice, document any adjustments or modifications, as well as any limitations of the data that have been identified.

Throughout the data-management process, visualisation is important for understanding and communicating what transformations may be required and the impact they have (visualisation techniques are discussed in more detail later).

In our example experience analysis, we used visualisations and summary statistics to explore the data, and used our actuarial and domain knowledge to validate the data before proceeding further. *Table 1* below compares actual deaths incurred with the PA(90) expected mortality.

Once exploratory analysis is complete, feature engineering takes place to process the data prior to modelling. In our example, we performed the following feature engineering techniques:

- Scaled features towards a normal distribution using a log transformation
- Encoded categorical features using one-hot-encoding
- Filtered the data by reducing the age range under observation to avoid fitting noise.

Figure 2: *Example comparison of actual and expected mortality by sex.*

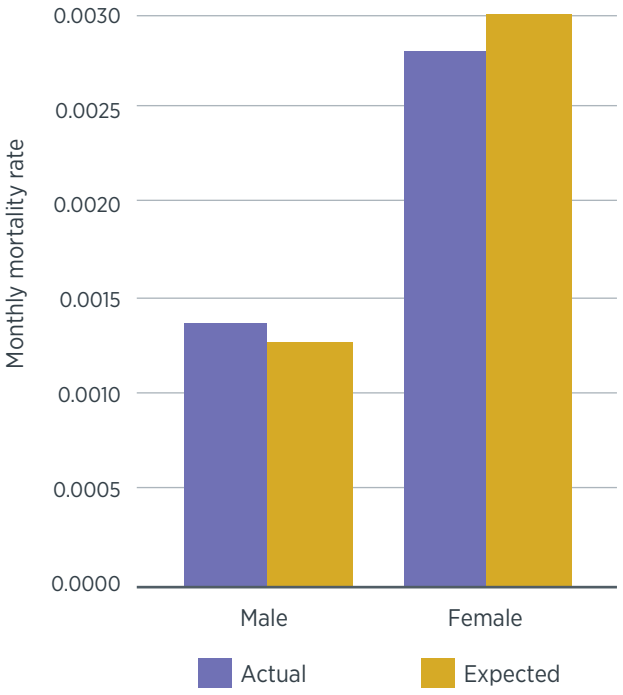


Table 1: *A summary A/E table showing in-force lives.*

Period	Actual mortality (deaths p/1000)	Expected mortality (deaths p/1000)	Actual/expected
1	1.94	2.70	72.1%
2	3.27	2.72	120.5%
3	2.71	2.66	101.8%

4. Model building

4.1 Training and testing

Model building is an iterative process in which techniques such as model fitting and model prediction using different algorithms are applied.

Before we fitted our model, we split the data as follows:

- **Training set** comprised 80% of credible data (post-transformations and adjustments) from the past three years, excluding the most recent year
- **Testing set** comprised the remaining 20% of credible data.

We could also vary the proportions as required. Note the most recent year's data was used as a validation set.

In this example our iterative model-fitting approach included:

1. A standard actual over expected (A/E) model was assumed (an A/E GLM of the Poisson family since mortality is assumed to increase exponentially with age) and fitted to the training set for calibrating the model
2. This model was tested for accuracy against data not included in the training set (20% out-of-sample), for instance, using the root mean square error (RMSE) or mean absolute error (MAE) to assess goodness of fit
3. The predicted rates and crude mortality rates were plotted against age as a curve to compare smoothness relative to accuracy
4. Based on the model performance and various standard model selection criteria such as Akaike information criterion (AIC), p-values, and domain knowledge, features are added or removed, and the model is retrained, retested and replotted
5. This process was repeated until we reached a satisfactory model
6. We then employed steps 1–5 again, using more complex models (a Poisson regression model applying Lasso regularisation),[2] until a reasonable set of candidate models were produced
7. The final model was selected from the shortlist generated by considering criteria such as prediction error, statistical measures, explainability (see **Section 5.1**), how close it was to the existing expected mortality model, and runtime, among others.

After performing the above tasks, we determined the following models were potential suitable predictive models for our example (see *Table 2* below):

- Lasso Poisson Regression Model (performed very similarly to a standard fitted Poisson model) – selected based on accuracy, runtime, and avoidance of overfitting
- Bayes[3] Poisson Regression Model – selected based on accuracy, ability to train on smaller samples of data, and ability for the model developer to specify prior distributions.

Lastly, we noted any limitations of our model, including age ranges where there were limited data points, as we wanted to ensure proper communication of the data and model limitations.

Mortality experience analysis often lends itself to regression analysis through GLMs, but alternative models such as generalised additive models (GAMs) and extreme gradient boosting machines (XGBoost) may be used to allow for irregularities and 'humps' in the data, to the extent that a closer fit is preferred over smoothness. In the table below, we list useful software packages for fitting GLMs as well as GAMs and GBMs, for example.

Table 2: *Comparing the candidate models against the actual experience and expected bases.*

	Expected basis	Lasso	Bayes
Mean Absolute Error (deaths p/1000)	3.86	3.92	3.62
Actual/Expected (%)	94.2%	93.1%	103.5%

2 | This is a method to penalise overly complex models and improve predictive ability across different datasets (Ng and Reid, 2021).

3 | Bayesian GLM modelling differs from traditional GLM modelling in that parameters are returned as distributions (in the form of posterior distributions), rather than point estimates (Barber, 2012). This allows for more flexibility in the modelling process (for examples, parameters belonging to different density functions), and the ability to stress-test assumptions by sampling different estimates from the parameters' distributions. In addition, since priors are provided for the distribution of parameters, expert judgement can be implemented where data is otherwise scarce.

Table 3: *Commonly used software packages for modelling.*

Algorithm	Example software package		
	Python	R	Julia
GLM	Statsmodels & Sci-kit Learn	Included in baseline R	GLM.jl
Lasso GLM	Statsmodels & Sci-kit Learn	glmnet	Lasso.jl
Bayesian GLM	PyMC3	rstan	Turing.jl
GAM	pyGAM	mgcv	At time of writing, there was no specific Julia-native package; it calls R's mgcv package
XGBoost	xgboost	xgboost	XGBoost.jl

4.2 Validation

A model that appears to perform well in one dimension (eg age) may perform poorly in another (eg socio-economic class) so we should analyse our model results from different angles. When validating a model, it is also important to consider model accuracy, stability and runtime, among others.

Care is needed to avoid selecting a model that fits the training sample too closely, as this could lead to poor predictive performance (due to overfitting). One technique for managing this issue is to use various training and validation samples when fitting and testing the model. If the model chosen shows a tendency to overfit to the training data, it may suggest that an alternative model should be considered.

Runtime is an important factor if the model will be used frequently. Typically, more accurate models require longer runtimes, so a balance is required between the accuracy preferred for the specific use and the time available for performing the analysis.

In our example, we considered:

- Variance in the predictive performance when filtering over a particular feature, such as sex
- Limitations of the model (eg at what years does it no longer produce reasonable results?)
- The run time of the model (eg it may be the case that a model has a better accuracy than the other but has a higher runtime than the other model)
- The stability of the model and other relevant factors (eg how well the model performs when given completely new data, such as the most recent years' set of data).

5. Reporting

5.1 Explainability

When considering model explainability, we refer to a process that aims to showcase the features that are most significant in the model's decision-making process. Techniques such as those listed below can help us understand and communicate models that could otherwise be perceived as black boxes.

- **Inherently interpretable models**, also white boxes, for example statistical models (eg naïve Bayes), linear models (eg linear and logistic regression), or additive models (eg Lasso).
- **Ex-post interpretable models** are potential black boxes (ie highly complex and where the relationship between inputs and outputs is not clear) where explainability techniques can be applied to help interpret and explain model results. Model-agnostic techniques include:
  - Variable importance plots: these show the average importance of each feature, in the sense of the extent to which the feature affects the target variable
  - Partial dependence plots: show the marginal effect of specific variables (usually just one or two) on the predicted outcome. This can show whether the relationship between the chosen feature and the outcome is linear, monotonic or more complex, for example
  - SHapley Additive exPlanations (Shapley values) show to what extent each variable contributed to the prediction for a single line in the dataset (Lundberg and Lee, 2017)
  - Local interpretable model-agnostic explanations (LIME) fit a simpler surrogate model that produces coefficients that indicate the impact different features have on a prediction (Ribeiro, Singh and Guestrin, 2016).

5.2 Visualisations and dashboarding considerations

Visualisations and dashboarding form an integral part of reporting and play an important role throughout our analysis, as this helps explain the process and results.

When reporting results of the modelling process, it is important to consider the stakeholders' level of familiarity with the analysis undertaken. More technical audiences, for instance, may appreciate tables and visualisations detailing model diagnostics. Those looking to extract business intelligence across multiple parameters may benefit more from an interactive dashboard rather than a static slide deck. Further considerations include:

- Knowing the stakeholder requirements before settling on tools and software for reporting
- Comparing at least one other model, as well as previous investigations performed
- Avoiding only presenting single point values of model performance (eg overall mean squared error), and instead include results across multiple dimensions (eg model performance with respect to smokers aged 40–45 vs non-smokers).

In our example experience analysis, a written static report with key visualisations was appropriate for stakeholders, along with the full codebase, presented in Jupyter Notebooks. Key visualisations included plotting aggregated data across various dimensions (age, smoking status, underwriting year, etc.) and mortality rate (actual, fitted (ie the model result), and the expected (as determined by the prior basis)). This allowed us to assess goodness of fit across various dimensions.

Figure 3: *Comparing baseline expected mortality to the A/E Lasso Poisson regression model by age.*

6. Other important considerations

6.1 Professionalism

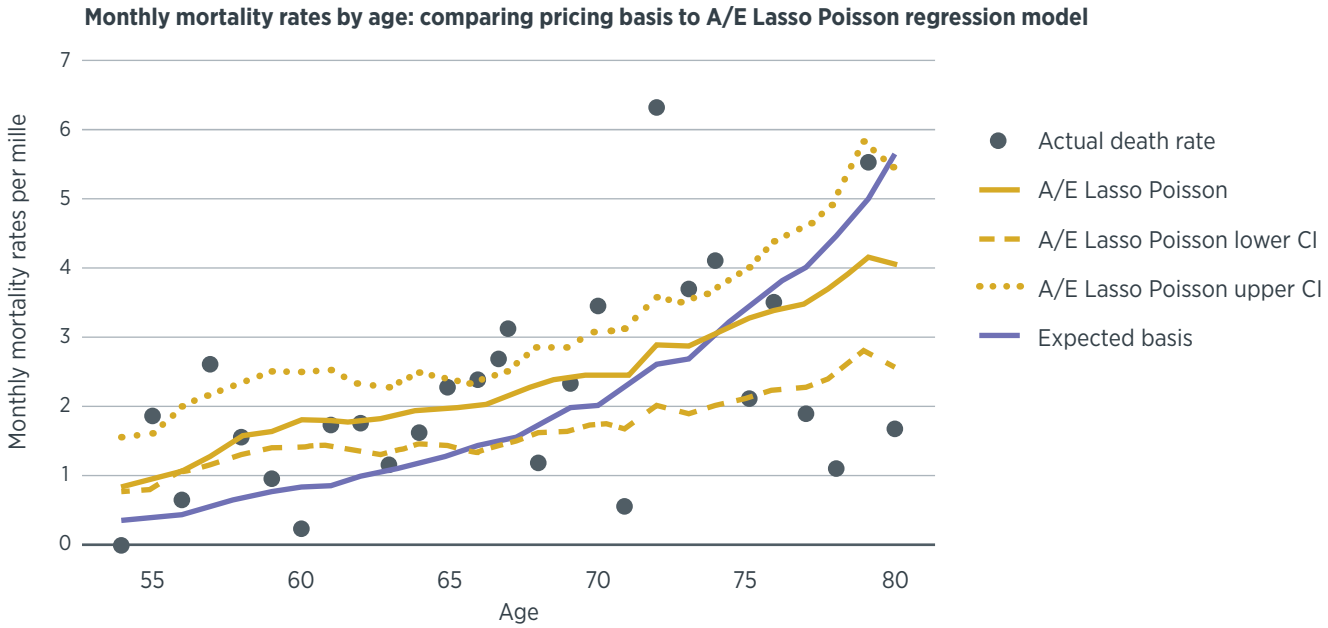
6.1.1 Risk management and governance

The actuarial solution framework should incorporate proper risk management, governance and control in all aspects of the pipeline.

Required considerations from governance and regulatory frameworks include those listed below, but note that other regulatory requirements, company specific policies or guidelines may apply.

- Integrity
- Compliance (regulatory and professional)
- Speaking up
- Judgement (reasonable and justifiable)
- Models (fit for purpose and sufficient controls)
- Communications (clear and comprehensive)
- Human oversight
- Data governance
- Transparency and explainability
- Fairness and non-discrimination
- Documentation and record keeping
- Robustness and performance.

These were derived from the Actuaries' Code (IFoA, 2019); TAS100 (FRC, 2023); EIOPA Guidelines for Insurance (EIOPA, 2021); and EU AI Act (Draft) (European Commission, 2021).



Furthermore, the key principles as set out in professional standards, such as APS X1 (IFoA, 2019b) and APS X2 (IFoA, 2015) should also be considered.

Other relevant regulations and guidelines may apply, for example Solvency II (European Parliament, 2009), GDPR (European Parliament, 2016) and FCA Fair Treatment of Customers (FCA, 2022) in the UK, as well as:

- Guidance from the Information Commissioner's Office (ICO) on:
  - AI and data protection (ICO, 2023)
  - Explaining decisions made with AI (ICO and The Alan Turing Institute, 2022)
  - AI auditing framework (draft guidance for consultation) (ICO, 2020)
- Any other guidance or legislation, for example from the Prudential Regulation Authority (PRA) in the UK
- Guidance from the AI Standards Hub in the UK, which provides a useful overview of relevant regulation across various industries and is working with the UK government to help inform a pro-innovation approach to AI assurance.

The guidelines and legislation above could form a key part of the governance and control framework for data-science related work. For further and more detailed information on how to interpret professional guidance, refer to the IFoA's Guidance for Members on Ethical and Professional Data Science (IFoA, 2021) and the IFoA and Royal Statistical Society's Guide to Ethical Data Science (IFoA and RSS, 2019).

The rate of development of AI tools and the availability of large language models (eg ChatGPT) present additional risks and require associated mitigating measures to be put in place. This article does not cover this, as we have focused on the core concepts of applying data science. However, an example issue that could arise is the ability of new models and AI tools to generate code which can be used to fit the models discussed above, amongst other uses. This raises questions regarding accountability, responsibility for outputs, and meeting professionalism requirements. Further research regarding an actuarial perspective is available on request.

6.1.2 Transparency

Transparency refers to the disclosure of information to stakeholders to understand the process a system or model followed, with relation to how the data is used by the model (and the solution overall), sources of external data (eg postcode socio-demographic classification indices, or credit ratings), the workings of the model, and in what context the outcomes will be used. In our example, technical documentation has been maintained considering the underlying data, assumptions, and approach to feature engineering and model development.

6.2 Monitoring and versioning

To ensure that changes to the analysis in all elements of the actuarial solution framework are committed and communicated sufficiently, we employ version control. After the model is implemented as part of the solution, the performance should be monitored and changes to the model or solution documented. Good version control practices and detailed documentation help developers and practitioners provide quality assurance. In order to collaborate in a secure, backed-up manner, tools such as GitHub and GitLab can be utilised. These tools can also assist in issue tracking and release management by allowing for a centralised platform.

Lastly, after the model is implemented as part of the solution, the performance should also be monitored on an ongoing basis.

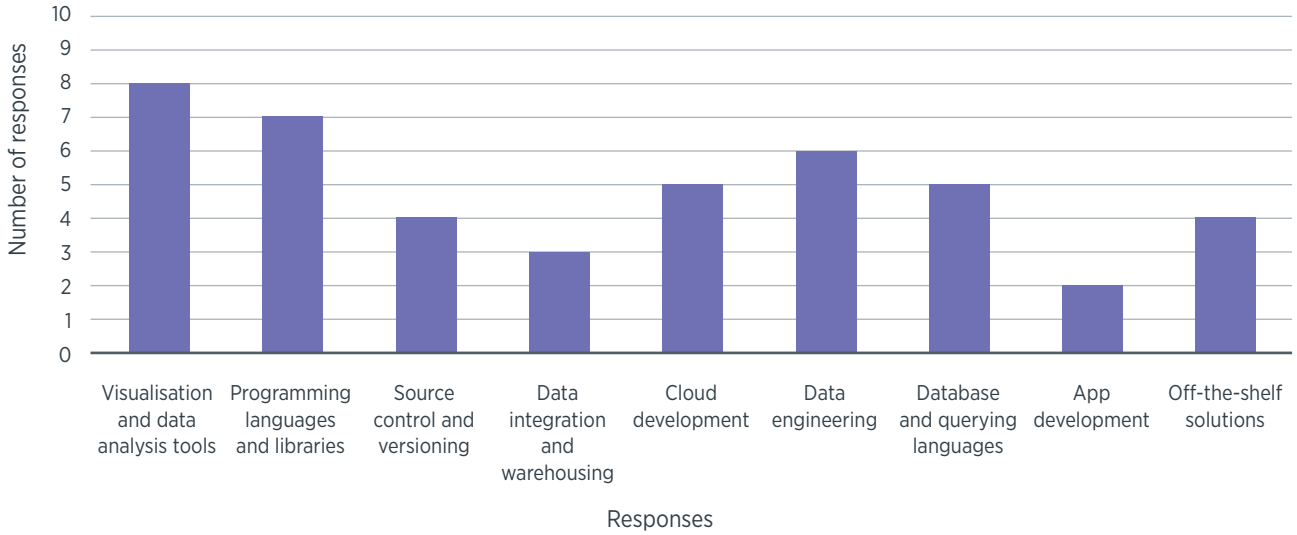
6.3 IT and infrastructure

IT and infrastructure decisions may be impacted by:

- Frequency of analysis:
  - How often will an analysis on new data be required?
  - How frequently, if at all, will models need to be retrained, or the whole experience analysis redone (ie rerun through all the steps in the data science pipeline)?
  - To what extent will models be monitored for drift and changes in underlying data?
- Data storage: How is our data stored, accessed and version controlled?
- Tools: What tools, software and/or platforms are utilised during our investigation eg how are we versioning our code base, how are we evidencing validations?
- Organisational infrastructure:
  - Do we have access to the required software, and the permissions required to perform our investigation? Is the available infrastructure fit for purpose?
  - Is the underlying infrastructure sufficient to support the implementation of the experience analysis eg the required computing power to perform the investigation, or the capacity to have a live dashboard that is constantly available to stakeholders? Is the environment secure enough to host our data and code?
  - Are tools and software utilised appropriately documented for handover to other teams, such as IT, internal audit, etc.?
  - What relevant IT and infrastructure policies are we required to comply with?

From *Figure 4* below we can see the use of various tools and techniques by actuarial teams that will influence the IT and infrastructure environment.

Figure 4: *Categories of tools and software used for analysis by actuarial teams (Actuartech, 2023).*



Additionally, by collaborating with other subject matter experts including data engineers, IT and data scientists, actuarial teams can ensure solutions and analysis produced can be used appropriately by other members of their organisation.

Conclusion

Data science and modern analytical techniques add value to traditional actuarial work by helping automate manual tasks, manage data efficiently, make improved predictions, and enhance decision-making. However, these techniques also need to be adopted ethically and securely, and used fairly.

References

Actuartech, Reacfin, and Synpulse (2021). *Beyond theoretical data science: a benchmarking of actuarial departments (White Paper)*. **https://www.actuartech.com/assets** [Accessed 28 July 2023.]

Actuartech (2023). *Superpowered actuaries: why incorporating modern IT & development skills could give you the edge. Industry Research on the actuary's role in infrastructure solutions.* **https://www.actuartech.com/assets** [Accessed 28 July 2023.]

Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge: Cambridge University Press.

Bellis, C., Lyon, R., Klugman, S. and Shepherd, J. (eds). (2010). *Understanding actuarial management: the actuarial control cycle* (2nd ed.). Sydney: Institute of Actuaries of Australia.

Dobson, A.J. and Barnett, A.G. (2008). *Introduction to Generalized Linear Models* (3rd ed.). Boca Raton: Chapman and Hall/CRC.

European Commission (2021). *Proposal for a regulation of the European Parliament and of the Council: laying harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts.* COM(2021) 206 final. **https://artificialintelligenceact.eu/the-act/** [Accessed 17 July 2023.]

European Insurance and Occupational Pensions Authority (EIOPA) (2021). *Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the European insurance sector.* **https://www.eiopa.europa.eu/publications/artificial-intelligence-governance-principles-towards-ethical-and-trustworthy-artificial_en** [Accessed 17 July 2023.]

European Parliament and the Council of the European Union (2009). *Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II) (recast)*. **https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009L0138** [Accessed July 17 2023.]

European Parliament and the Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. **https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679** [Accessed July 17 2023.]

Financial Conduct Authority (FCA) (2022). *Principles for good regulation.* **https://www.fca.org.uk/about/how-we-regulate/handbook/principles-good-regulation** [Accessed 17 July 2023.]

Financial Reporting Council (FRC) (2023). *Technical Actuarial Standard 100: general actuarial standards.* **https://www.frc.org.uk/actuaries/technical-actuarial-standards** [Accessed 17 July 2023.]

Information Commissioner's Office (ICO) (2020). *ICO consultation on the draft AI auditing framework guidance for organisations.* **https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/** [Accessed July 17 2023.]

Information Commissioner's Office (ICO) (2023). *Guidance on AI and data protection.* **https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/** [Accessed July 17 2023.]

Information Commissioner's Office (ICO) and The Alan Turing Institute (2022). *Explaining decisions made with AI.* **https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/** [Accessed July 17 2023.]

Institute and Faculty of Actuaries (IFoA) (1990a). *Life Office Pensioners, females, Normals, amounts.* **https://www.actuaries.org.uk/learn-and-develop/continuous-mortality-investigation/cmi-mortality-and-morbidity-tables/mortality-rates-older-mortality-tables** [Accessed 17 July 2023.]

Institute and Faculty of Actuaries (IFoA) (1990b). *Life Office Pensioners, males, Normals, amounts.* **https://www.actuaries.org.uk/learn-and-develop/continuous-mortality-investigation/cmi-mortality-and-morbidity-tables/mortality-rates-older-mortality-tables** [Accessed 17 July 2023.]

Institute and Faculty of Actuaries (IFoA) (2015). APS X2: Review of actuarial work. **https://actuaries.org.uk/standards/work-review** [Accessed 17 July 2023.]

Institute and Faculty of Actuaries (IFoA) (2019a). *The actuaries' code.* **https://actuaries.org.uk/the-actuaries-code/** [Accessed 17 July 2023.]

Institute and Faculty of Actuaries (IFoA) (2019b). APS X1: Applying standards to actuarial work, in actuarial Professional Standards: **https://actuaries.org.uk/standards/standards-and-guidance/actuarial-profession-standards-aps/** [Accessed 17 July 2023.]

Institute and Faculty of Actuaries (IFoA) (2021). *Ethical and professional guidance on data science: a guide for members.* **https://actuaries.org.uk/standards/data-science-ethics/** [Accessed July 17 2023.]

Institute and Faculty of Actuaries (IFoA) and Royal Statistical Society (RSS) (2019). *A guide for ethical data science.* **https://actuaries.org.uk/standards/standards-and-guidance/non-mandatory-guidance/** [Accessed 17 July 2023.]

Lundberg, S.M. and Lee, S.-I.. (2017). *A unified approach to interpreting model predictions.* **https://doi.org/10.48550/arXiv.1705.07874**

Ng, J. and Reid, S. (2021). *Mortality impact of COVID-19 vaccination in England: counterfactual insights from Gompertz to machine learning.* **https://www.actuaries.org.uk/news-and-insights/news/mortality-impact-covid-19-vaccination-england** [Accessed 17 July 2023.]

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). *"Why should I trust you?": explaining the predictions of any classifier.* **https://doi.org/10.48550/arXiv.1602.04938**

Valerie du Preez, FIA

Valerie is a qualified actuary specialising in actuarial transformation and the founder of Actuartech, supporting clients with the training and implementation of data science and technology in a business context. She has been involved in a number of key industry roles to understand the value and risk of data science, including the IFoA's Data Science Working Party, the IFoA's Certificate in Data Science, and as a task force member working on the ethical guidance for practitioners working in data science. She leads a cross-regional research group on AI risk in the context of actuarial work and is a member of the FRC Advisory Panel. Contact: **info@actuartech.com**

Patrick Moehrke, TASSA

Patrick is leading corporate students on their data science and technology journey in R, Python and cloud-based services. He uses his knowledge of machine learning techniques, including his experience in R, Python and Julia, to help identify ways to solve business challenges, including investigating the impact of customer behaviour on policy cancellations, investigating the impact of Covid-19 on mortality and longevity experience, and investigating different modelling and explainability techniques to help manage AI risk.

Lara van Heerden

As business development lead at Actuartech, Lara leads the research and publication activities, and assists in managing the standard of training provided to ensure it addresses the client's needs. Recent research activities include an examination and discussion of the risks of AI and ML from an actuarial perspective, benchmarking the data science and data management practices of actuarial functions, and investigating the impacts of different regulatory and statutory requirements on the actuarial operating model

Estimating neighbourhood death rates using the random forest algorithm

Andrew Cairns, Jie Wen and Torsten Kleinow

Introduction – the longevity problem we address

Recent decades have seen increasing evidence for inequality in mortality for different socio-economic groups in various national populations. Socio-economic characteristics are not the actual cause of mortality inequalities. Rather, the socio-economic characteristics of a population are correlated with the prevalence of various health-related lifestyles such as smoking, diet and exercise.

In addition, they can be related to the availability of preventive health care, crime rates, air pollution and other external factors that have an impact on health and mortality. Often, observed inequalities are based on existing socio-economic indicators such as income (eg Chetty et al., 2016), affluence (Cairns et al., 2019), deprivation (Villegas and Haberman, 2014) or education (eg Mackenbach et al., 2003, 2015).

However, many of these metrics are designed for other purposes. This means that while the English Index of Multiple Deprivation, for example, can be used as a good predictor of mortality, perhaps we can do better by designing a customised mortality index. Specifically, it might be possible to improve on these existing approaches by combining individual pieces of socio-economic information at the individual or (as we do here) neighbourhood level and analyse this using modern data-science techniques: here, the random forest algorithm. This method will allow us to capture how mortality rates respond to a range of variables, potentially in a non-linear way.

The structure of the model

In general, our data consists of a set of observations. Each observation has a set of predictive variables and a single response variable. In our case study:

- The number of observations represents the number of neighbourhoods (Lower Layer Super Output Areas or LSOAs)
- The predictive variables are socio-economic and related variables that give an indication of the general character of each LSOA
- The response variable is the ratio of actual deaths in a specific LSOA over a specified range of ages and years relative to what would be expected if the LSOA had the same base mortality rate as the national population.

Our challenge is to predict the actual-to-expected death ratio (which we refer to as the relative risk) as a function of the predictive socio-economic variables. So, effectively, we have a regression problem to solve. We do so by making use of the random forest algorithm.

This is a widely used machine learning algorithm that combines the output of multiple regression trees (also known as decision trees) to determine a single result. It can be used for classification and regression problems and the flexibility that it offers is a key reason for its popularity. In our case, we will use this method to find an 'average prediction model', using a selection of samples of subsets of the data, randomly chosen, and capturing varying degrees of features of the predictive variables. Using multiple trees has additional advantages of flexibility, including greater accuracy, a reduced risk of overfitting, and the ability to determine the importance of each variable in the model (using a chosen metric). This comes at the cost of greater computing times and complexity of the model, which can sometimes make interpretation of the results more difficult.

Regression trees

Regression trees themselves are a form of supervised machine learning, where a sequential process is followed to split the data of interest. In our case, this is the relative risk response variable. Each tree consists of non-overlapping nodes (sometimes called leaves), splits and predictors. Moving through the tree can be considered as answering a series of questions, such as 'Is the value of a specific predictive variable greater than 0.3?'. Depending on the answer to this question, further subsequent questions might be 'Is the value of the predictive variable greater than -0.6 (if the answer to the

original question was no)?'. When the optimal split has been determined, we arrive at a node in the tree. In our example, each node will be a value of the predictor function for the response variable, which is estimated as the mean of the subset of the observations of the response variable that is allocated to the specific node. This is shown in *Figures 1* and *2* below.

In this stylised example, illustrated in *Figures 1* and *2*, we have two predictive variables and 1,000 observations There are currently three splits and the whole of the area is covered by the four nodes. The splits are numbered in the order they took place. Splits 1 and 2 used predictive variable X_1 to divide the data, first at X_1 = 0.3 then at X_1 = -0.6. Split 3 then found that it was optimal to use predictive variable X_2 to split the righthand node above and below X_2 =0.6. The general algorithm works in a similar way but considers a larger number of predictive variables and makes more than three splits.

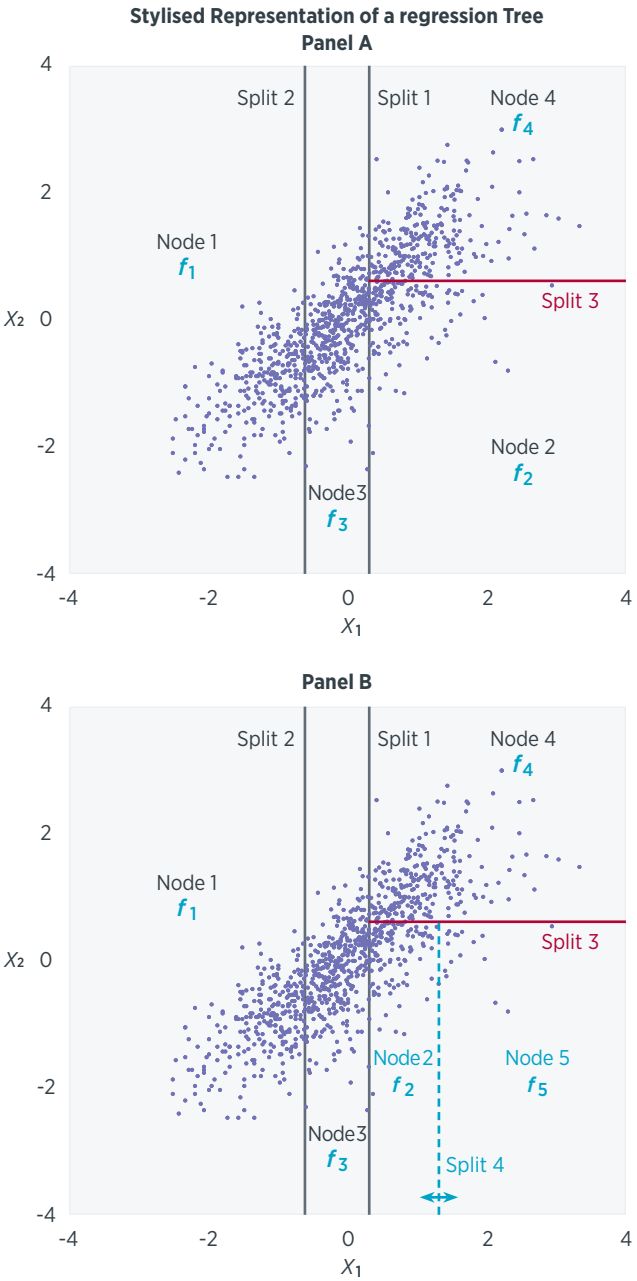


Figure 1: *Stylised representation of how a regression tree grows with two correlated predictive variables X_1 and X_2. Panel A: tree consists of nodes, splits and function values. Panel B: split 4 optimises over the location of the split and the updated function values either side of the split.*

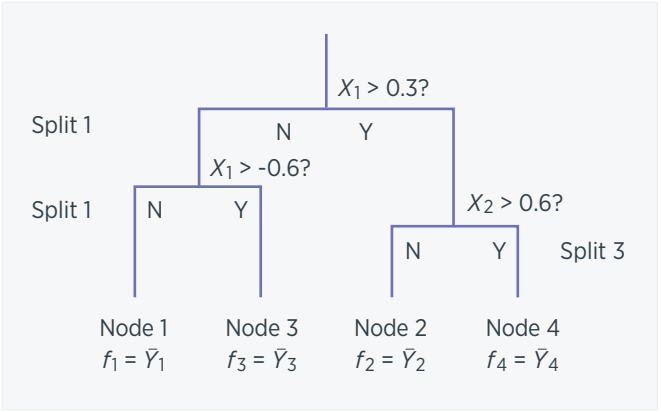


Figure 2: *Graphical representation of the regression tree corresponding to the left-hand panel of Figure 1. Y_k is the mean of the observations falling within Node k.*

The number of splits made in the tree (equivalently the number of questions asked) will depend on a selection criterion. We have chosen the residual sum of squares (RSS), aiming to minimise this when considering the difference between the combined predictor for the response variable and the actual observations that we have of that variable.

Increasing the number of splits in the tree

Splits are added by scanning over each node, each predictive variable within each node, and the position of the split of that predictive variable within that node.

To choose the fourth split in our stylised example we optimise along the following lines. Minimise the RSS over:

- Each of the pre-existing nodes (eg node 2 in *Figure 1*, Panel A) (ie only a single node is split into two rectangles; all other nodes remain as they were)
- Each of the two predictive variables (eg a vertical or horizontal split of node 2 in Panel A)
- The position of the split horizontally or vertically (eg the vertical split 4 in Panel B)
- (Assuming we are considering a split of the pre-existing node 2 into a reduced node 2 and new node 5) the values f_2 (updated from the old f_2) and f_5 taken by the function either side of the new split. The optimal values for the function $f_{(x)}$ at any point in the newly divided nodes (nodes 2 and 5 in Panel B) are the mean of the observations within each of the two new nodes.

Splitting stops when the algorithm reaches a specified stopping criterion. The idea is to add increasing granularity to the piecewise constant estimator of the response, in order to make the combined estimator (ie considered across all of the data simultaneously) closer to the observed results, while avoiding overfitting.

A detailed description of the algorithm and its application in a mortality context can be found in Wen et al. (2023). Here, we provide an overview of its main elements. For further detail, including treatment of categorical variables, see James et al. (2021) or Hastie et al. (2017).

The random forest algorithm

It is often found that individual regression trees can suffer from relatively high levels of uncertainty. For example, if we split the data into two randomly chosen halves and fit a regression tree to each, the results can be quite different.

The random forest algorithm (RF) is a popular approach that substantially reduces this uncertainty. RF works by combining the results for a defined number of individual trees, with the uncertainty reduction depending on differences in how each tree is grown.

The original observations are first divided into two parts: a training dataset and a validation dataset.

For each tree grown (by using the splitting process described above):

- The tree is fitted to a randomly chosen subset of the observations in the training dataset. It is common (see, e.g., James et al., 2021) for the subset to be formed by taking a random sample with replacement from the training dataset. This results in a sample which contains approximately 2/3 of the training dataset.
- As each tree grows, for each split, rather than optimise over all possible predictive variables, we optimise over a randomly selected subset of predictive variables. The size of this subset is usually substantially less than the total number of predictive variables
- Each tree keeps growing so long as all nodes contain at least a specified number of observations
- Each tree produces an estimator and the random forest estimator is then the arithmetic average of all predictors for all of the individual regression trees.

All the training datasets are drawn from the full training subset of the full set of observations. To validate the model, the remaining observations are then used as an 'out-of-bag' sample to test the accuracy of the estimator. The out-of-bag sample also allows us to tune the choices for the number of trees, the size of the subset of predictive variables and the minimum number of observations required for each node in the tree.

Once validation is complete and the choices for the above variables (referred to as 'hyperparameters') are fixed, the model can then be rerun on a different (but still randomly selected) training set and tested for goodness of fit on the complementary `test' dataset. (This training/test division can also be used to compare the outputs of the random forest with alternative models.) If all of this is satisfactory then the final step is to run the random forest algorithm on the full set of data.

Standardisation of predictive variables

In some settings the predictive variables might be standardised in some way (for example, by transforming the data so that it has a standard normal distribution). In the context of the RF algorithm, the purpose of this is mainly to facilitate graphical analysis of the results (see, for example, ***Figure 5***). Provided a transformation of the data preserves the order of the observed values, the results of the RF algorithm are not sensitive to such transformations. This contrasts with, for example, generalised linear models: if a relationship is linear, then anything other than a linear transformation of the data results in a qualitatively different model. This lack of sensitivity is a key advantage of the RF algorithm. It means the user does not have to spend time thinking about how to standardise (or otherwise) the data and allows them more time to think about other issues.

Application to English neighbourhood mortality

As noted earlier, our data consists of deaths and exposures for single ages from 2001 to 2018 with twelve predictive variables for each LSOA. Please note that we have not applied age standardisation when comparing actual and expected deaths. The predictive variables used in this study (after much time spent selecting variables) were:

- Old-age income deprivation
- Employment deprivation
- Proportion above age 65 with no qualifications
- Crime rate
- Average number of bedrooms
- Proportion born in the UK
- Wider barriers to housing (eg homelessness and affordability)
- Proportion in management positions
- Proportion working more than 49 hours per week
- Urban-rural class
- Proportion aged 60+ in a care home with nursing
- Proportion aged 60+ in a care home without nursing.

Of these, the deprivation measures relating to old-age income deprivation and employment deprivation are the most important drivers of mortality inequality. Also important at the neighbourhood level are urban-rural class and proportions in a care home, both with and without nursing. The remaining predictive variables are less important but have been found to be statistically significant, sometimes in unexpected ways (see the discussion of ***Figure 5***).

As we discuss below, the urban-rural class turns out to be an important predictive variable that is missing (at least explicitly) from the alternative Index of Multiple Deprivation (IMD) (Ministry of Housing, Communities and Local Government, 2015). The five urban-rural classes are:

1. Urban conurbations excluding London
2. Urban cities and towns
3. Rural towns and villages
4. Rural hamlets and isolated dwellings
5. Urban conurbation in London only.

The inclusion of nursing home proportions allows us to adjust for the distorting effect on the mortality of neighbourhoods with care homes. In doing so, the resulting Longevity Index for England (LIFE) gives a measure of mortality that reflects the mortality of the non-care home population in an individual LSOA relative to the national average. For further details, see Wen et al. (2023).

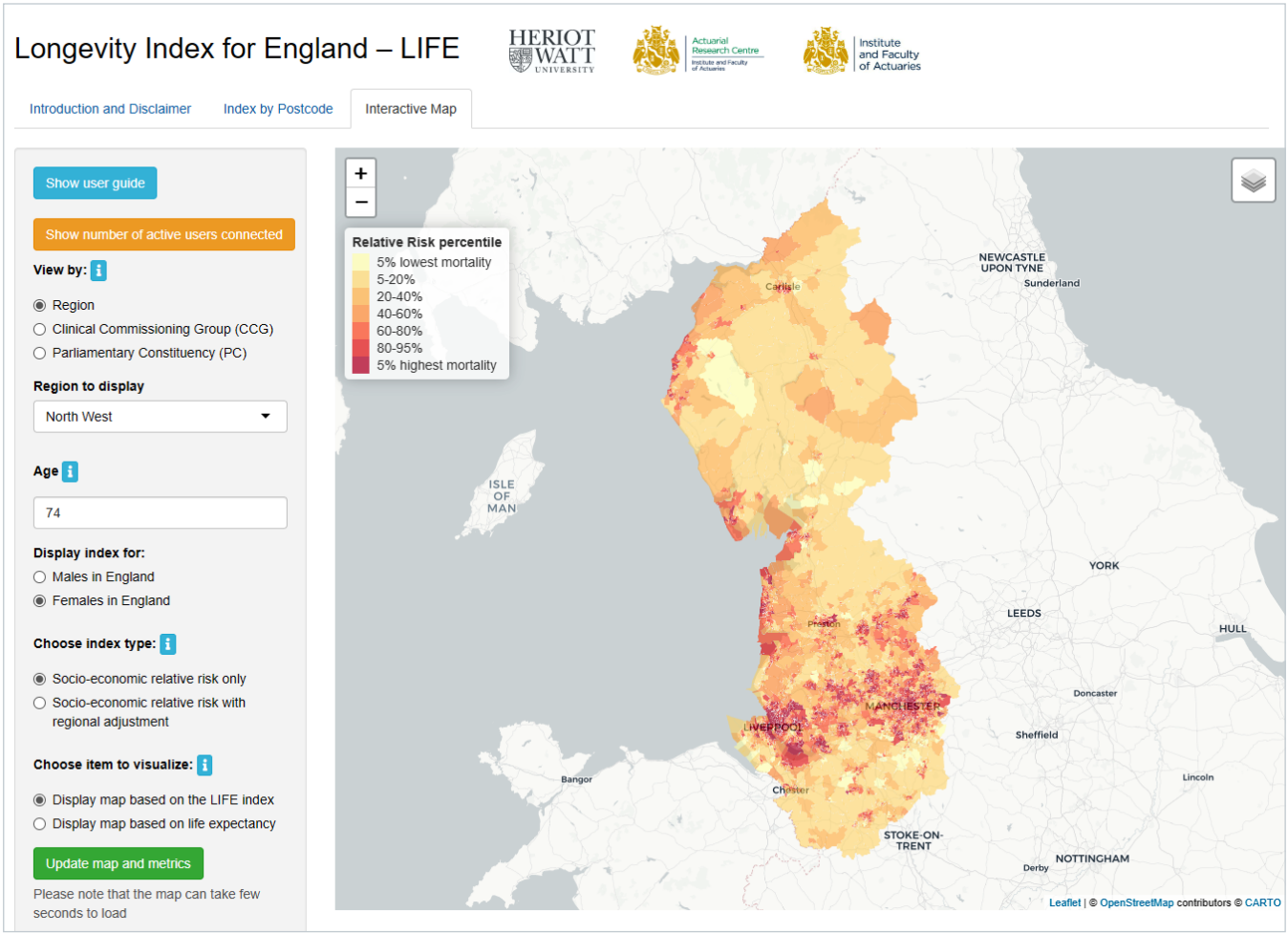
Here, we present a small number of snapshots of what is a complex set of inputs and outputs and potential graphics.

The LIFE app

Readers can explore the results in more detail with the open-source LIFE app that can be found at **https://andrewcairns.shinyapps.io/LIFEapp_Version3/**. The app allows users to look at individual LSOAs, males and females, and different ages. It also allows users to map the LIFE index to observe how the LIFE index varies throughout a region.

An example of this is shown in *Figure 3* for the North West of England. The map illustrates clearly how areas of high mortality are concentrated in the larger cities, an observation that is at least partly explained by concentrations of high deprivation in these areas.

Figure 3: *Example of a screen shot from the LIFE app showing relative risk percentiles for females aged 74 in the North West of England.*



The impact of urban-rural class

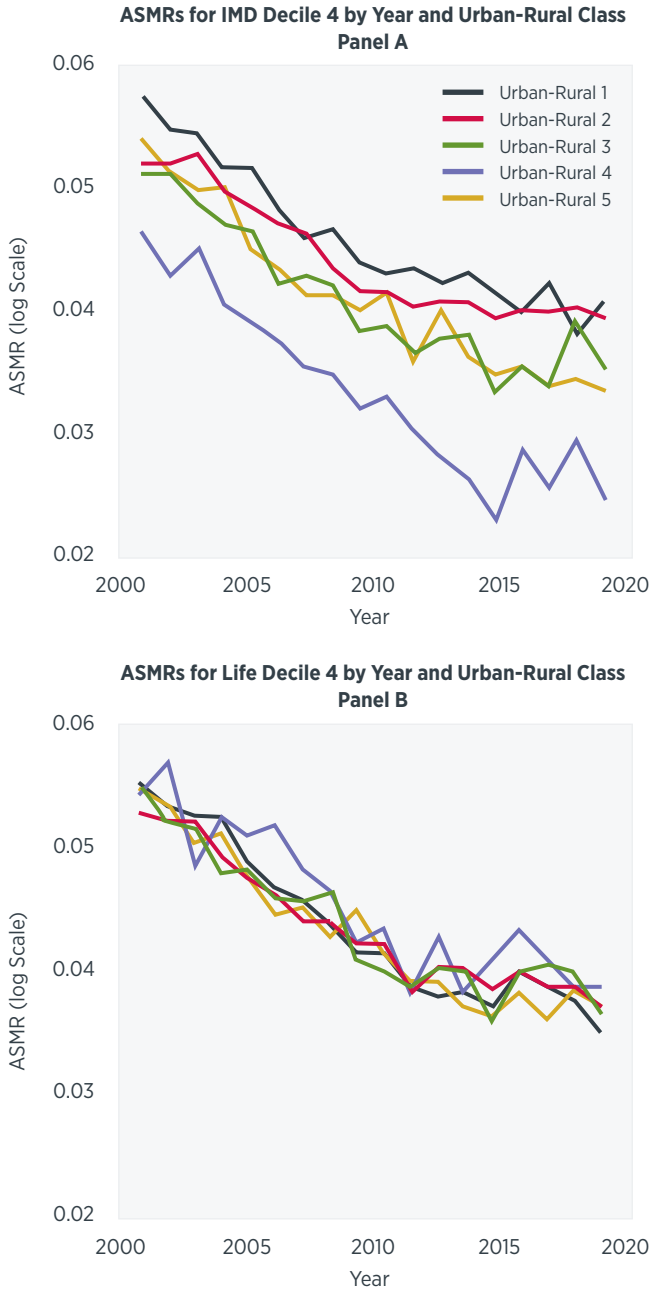


Figure 4: *ASMR for ages 70–79 for Decile 4 from 2001–2018 for the five urban-rural classes. Panel A: deciles based on the IMD. Panel B: deciles based on the RF LIFE index.*

In *Figure 4* we highlight how poor the IMD is at discriminating between urban and rural areas in terms of predicting mortality. In the left-hand panel we show how the age-standardised mortality rate (ASMR) varies over time for IMD decile 4 (by way of example) subdivided into the five urban-rural classes. That is, we only count deaths and exposures for LSOAs that belong to IMD decile 4 and a specific urban-rural class. If the IMD was a good predictor of mortality, then there should be little difference between the ASMRs for the five urban-rural classes. Instead, what we see for decile 4 is considerable variation between the urban-rural classes.

In particular, mortality in the most rural areas (class 4) is much lower. In contrast, decile 4 based on use of the LIFE index (right-hand panel) shows very little variation between the urban-rural classes, mainly as a result of incorporating urban-rural class as a predictive variable. This demonstrates that the LIFE index and the random forest algorithm has, at least, picked up this feature of the data.

Dependence on key predictive variables

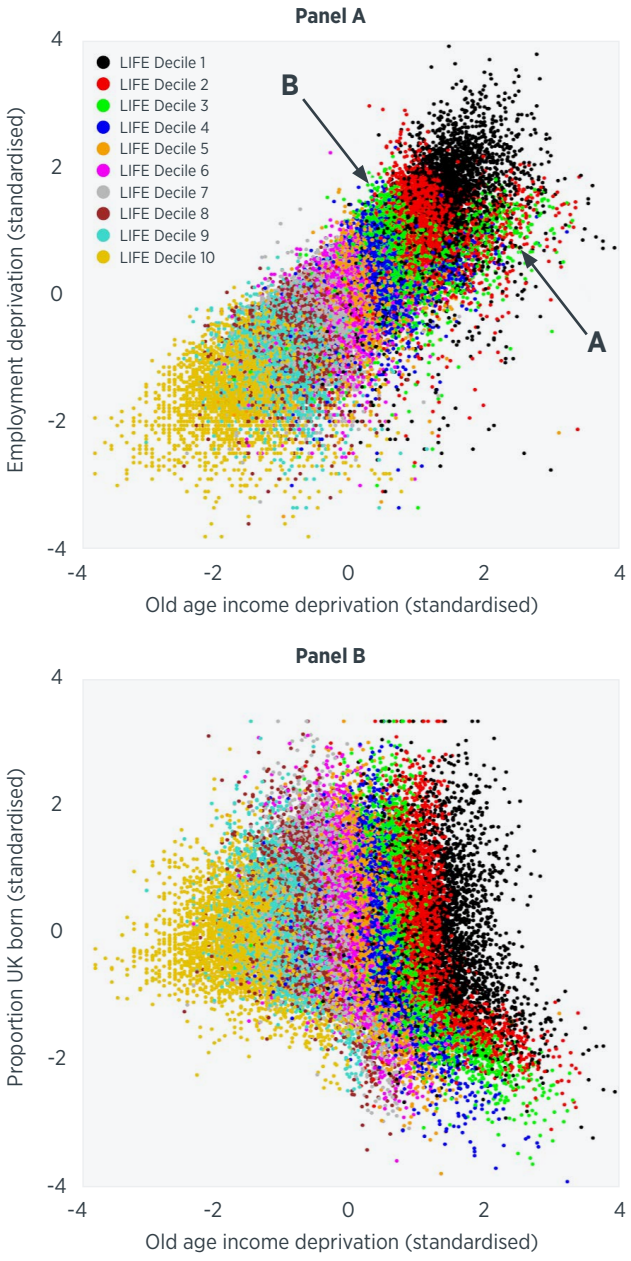


Figure 5: *Scatterplot showing how the LIFE decile for ages 70–79 depends upon two key variables. Panel A: old-age income deprivation and employment deprivation. Panel B: old-age income deprivation and the proportion of UK born.*

Further graphical investigations revealed that LSOAs that belonged to feature A were mainly LSOAs that had a low proportion of UK born. Thus, in the right-hand panel we show how the LIFE decile depends on old-age income deprivation and the proportion of UK born. The striking feature of this plot is how the characteristics in the lower part of the plot (when the standardised value of the proportion of UK born is below about -1) differ markedly from the upper portion. The upper portion suggests that the proportion of UK born has very little predictive power. In contrast, the lower portion (about 15% of the data, mostly in London and larger cities) suggests that the proportion of UK born is also as strong a predictor of mortality as old-age income deprivation.

From a statistical perspective, the results suggest that, on a like-for-like basis, neighbourhoods with a low proportion of UK-born people have lower mortality. It is less clear how we interpret this unusual feature. One possibility out of many is that deprived neighbourhoods with a high proportion of immigrants are good at looking after each other or follow healthier lifestyles. It is certainly a feature that merits further study.

The right-hand plot also highlights a big advantage of the random forest algorithm over traditional linear models: the non-parametric nature of the RF algorithm means that it can easily pick up unusual and localised features of the data.

Regional mortality

In Table 1 we investigate how well the LIFE index performs as a predictor of mortality at the regional level. Each region shows actual over expected (A/E) deaths by region for males 70–79, 2001–2018. The unadjusted column shows the raw A/E based on national mortality only, and echoes the much-discussed North-South divide in mortality. The middle column, Adjusted IMD, adjusts expected deaths to account for variation in the distribution of deprivation based on the IMD by region. This reveals that much of the variation that we observe in the unadjusted A/E is due to variation in deprivation, but some variation remains. The final column adjusts using the LIFE index (including actual care home proportions) rather than the IMD, and we can see that regional variation has further significant reductions. So, again, we can see that the LIFE Index does a better job at explaining variation in mortality across England using socio-economic and related non-spatial predictive variables only.

However, from *Table 1*, we can see that the LIFE index does not do a perfect job: some regional differences remain. Understanding the remaining regional differences is beyond the scope of this article. But, in brief, cause of death data might help; for example, lung-cancer mortality by region and income deprivation clearly indicates that smoking prevalence by deprivation is significantly higher in the northern regions than in the south. In other words, while the prevalence of smoking, as a causal risk factor, has a very strong dependence on socio-economic status, there is additional variation by region. This, in turn, has an impact on all-cause mortality at the regional level that reveals itself in the final column of *Table 1*.

Table 1: *Actual versus expected deaths for males aged 70–79 over the period 2001–2018 by region. Unadjusted: expected deaths based on national mortality by single year and single age. Adjusted IMD: national mortality rates are multiplied by a relative risk based on IMD rankings. Adjusted LIFE: national mortality rates are multiplied by a relative risk based on the LIFE index.*

	Actual over Expected Deaths (%)		
	Unadjusted	Adjusted	
Region		IMD	LIFE
North East	115.5	106.1	101.1
North West	112.9	106.0	103.8
Yorkshire and the Humber	107.6	102.8	101.5
East Midlands	101.8	102.6	101.1
West Midlands	104.2	100.3	99.1
East	91.5	96.1	97.1
London	99.5	95.4	99.1
South East	90.4	99.9	100.3
South West	89.2	92.8	96.2

Summary and key findings

We have described how the random forest (RF) algorithm works and how it can be applied to English neighbourhood data to reveal new insights into how mortality and longevity varies across the country through use of the Longevity Index for England (LIFE). We find that the LIFE index delivers a significant improvement over the IMD as a predictor of neighbourhood mortality. In part, this is because the index is tailored to the modelling of mortality outcomes. But it is also because the flexibility of the random forest algorithm allows us to incorporate a greater variety of predictive variables with potentially non-linear effects and interactions. The resulting improved fit reveals the following key points:

- Old-age income deprivation and employment deprivation are the key predictors of mortality rates
- Urban-rural class (which is not a component of the IMD) along with proportions in a care home are also important as predictors
- Using IMD deciles gives a misleading impression that there is significant, unexplained variation in mortality at the regional level. The LIFE index demonstrates that much of the variation that we see at regional level can be explained by making proper allowance for socio-economic and urban-rural neighbourhood characteristics.

The RF algorithm does not assume a linear relationship between predictive variables and response variables: it just picks up whatever shape relates one to the other. Similarly, there is no need to specify in advance any interactions between combinations of predictive variables and the response variable: the algorithm detects and incorporates these automatically. This contrasts with more traditional methods of mortality modelling, such as generalised linear models (GLMs) where any interactions need to be individually identified and incorporated into the model. Nevertheless, GLMs do have the advantage of greater interpretability, although graphical analysis of RF inputs and outputs can help considerably with interpretation.

As we have only presented some of the empirical findings based on the LIFE index, we invite the reader to explore this topic further through the LIFE app mentioned above, or by contacting the authors.

References

Cairns, A.J.G., et al. (2019). Modelling socio-economic differences in the mortality of Danish males using a new affluence index. *ASTIN Bulletin* 49(3): 555-590. **https://doi.org/10.1017/asb.2019.14**

Chetty, R., et al. (2016). The association between income and life expectancy in the United States, 2001-2014. *JAMA* 315(16): 1750-1766. **https://doi.org/10.1001/jama.2016.4226**

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). New York: Springer. **https://hastie.su.domains/ElemStatLearn/** [Accessed 1 June 2023.]

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). New York: Springer. **https://www.statlearning.com** [Accessed 1 June 2023.]

Mackenbach, J.P., et al. (2003). Widening socioeconomic inequalities in mortality in six Western European countries. *International Journal of Epidemiology*, 32(5): 830-837. **https://doi.org/10.1093/ije/dyg209**

Mackenbach, J.P., et al. (2015). Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries. *Journal of Epidemiology & Community Health*, 69(3): 207-217. **https://doi.org/10.1136/jech-2014-204319**

Ministry of Housing, Communities and Local Government (2015). *English indices of deprivation 2015: research report.* **https://www.gov.uk/government/publications/english-indices-of-deprivation-2015-research-report** [Accessed 1 June 2023.]

Villegas, A.M., and Haberman, S. (2014). On the modeling and forecasting of socioeconomic mortality differentials: an application to deprivation and mortality in England. *North American Actuarial Journal*, 18(1): 168-93. **https://doi.org/10.1080/10920277.2013.866034**

Wen, J., Cairns, A.J.G., and Kleinow, T. (2023). Modelling socio-economic mortality at neighbourhood level. *ASTIN Bulletin*, 53(2): 285-310. **https://doi.org/10.1017/asb.2023.12**

Acknowledgements

This work forms part of the research programme 'Modelling, Measurement and Management of Longevity and Morbidity Risk' funded by the Actuarial Research Centre of the Institute and Faculty of Actuaries, the Society of Actuaries and the Canadian Institute of Actuaries.

The authors gratefully acknowledge feedback from the respective steering and oversight groups.

This study is also part of the research programme at the Research Centre for Longevity Risk – a joint initiative of NN Group and the University of Amsterdam, with additional funding from the Dutch government's Public Private Partnership programme.

Andrew Cairns

Andrew Cairns is Professor of Actuarial Mathematics at Heriot-Watt University, Edinburgh and at the Maxwell Institute for Mathematical Sciences.

He is well known both in the UK and internationally for his research in financial risk management for pension plans and life insurers. In recent years his research has focused on the modelling of longevity risk: how this can be modelled, measured and priced, and how it can be transferred to the financial markets. Among his work in this field, he has developed a number of new and innovative stochastic mortality models. He has also worked extensively on inequalities in cause of death mortality and the evolving impact of Covid-19 on current and future mortality.

He is an active member of the UK and international actuarial profession. He qualified as a Fellow of the Faculty of Actuaries in 1993 and from 1996 to 2017 was editor of the leading international actuarial journal *ASTIN Bulletin - The Journal of the International Actuarial Association*. In 2005 he was elected as a corresponding member of the Swiss Association of Actuaries. From 2016–2020 he was Director of the Actuarial Research Centre of the Institute and Faculty of Actuaries.

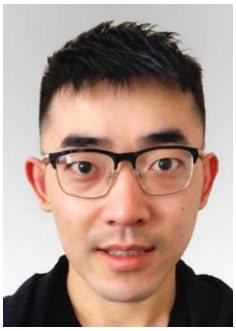
His research has received several international prizes including the Halmstad Prize in 2008, the Society of Actuaries Annual Prize in 2009 and the Robert I. Mehr Award in 2016.

In 2016 he was elected as a Fellow of the Royal Society of Edinburgh, Scotland's national academy of science and letters.

Torsten Kleinow

Torsten Kleinow is professor at the University of Amsterdam (UvA) and director of the Research Centre for Longevity Risk at the Amsterdam School of Economics. Before joining UvA he worked as associate professor at Heriot-Watt University in Edinburgh and was a member of the IFoAs Actuarial Research Centre. His research on mortality modelling and related actuarial topics has been published in leading academic journals and presented at many academic and industry events. He currently serves as an editor for the *European Actuarial Journal.*

Jie Wen



Jie Wen works at Lloyds Banking Group as a Pricing Manager. He has an MSc in Actuarial Management and a PhD in Actuarial Mathematics from Heriot-Watt University. His PhD research focused on the modelling of mortality inequalities and machine learning methods for modelling mortality rates

A history of actuarial engagement with electronic health records

Dan Ryan, Director of Demographic Risk at Just

The digitisation of health records has created significant opportunities for research, as well as improving the accessibility of medical records for the patients themselves. Information that in the past was only available in voluminous paper files can now be viewed, and even amended, using mobile phones. This can include diagnoses, observations, investigations and changes in medications.

This transformation has progressed at different speeds through the UK health system, with primary care leading the way. The coding of existing paper records in almost every GP practice was a gargantuan task, but it has supercharged the patient consultation. GP practices have been capturing terabytes of data onto primary care software systems such as EMIS Health, SystmOne, iSORT and INPS Vision. Rapid exchange of medical information, automated prescribing, and a more insightful picture of the patient's health are now possible.

This data transformation allowed large patient databases such as the General Practice Research Database (now Clinical Practice Research Datalink (CPRD)) and The Health Improvement Network (THIN) to be established, and made anonymised electronic health records (EHR) available for research. Access to each EHR dataset requires ethical approval and justification of the public health benefits of any research protocol. Worldwide academic institutions and the pharmaceutical industry were the most prolific investigators. Indeed, it was only through these anonymised EHR datasets that the pharmaceutical industry had clear sight over how, where and when medications were being used and with what outcomes.

Actuaries and the insurance industry were much more cautious about the benefits of EHR datasets, put off by a combination of licensing costs of £300,000+ per annum, doubts over the veracity and completeness of the underlying data, and long-standing concerns over the applicability of general population mortality and morbidity experience to insured portfolios and pension schemes. Part of this reticence likely reflected the expectation that it would take significant time and effort to demonstrate to regulators that the innovative uses of such datasets would not only improve future best estimates (and hence reduce premiums) but also reduce uncertainty in future assumptions and recognise this through lower requirements in risk capital.

Nevertheless, pioneers across the actuarial profession took the first tentative steps in the early 2000s, analysing anonymised data from GPRD to support new mortality models. Publications soon followed, such as the SIAS publication *Disease and Death* by Love and Ryan in 2007. But even now, the number of insurers and reinsurers involved in research into EHR datasets is limited. Indeed, in the decade before Covid, public and political concerns meant that access for insurers and reinsurers became harder, but was still possible through financial support of research projects at academic institutions.

One research collaboration between Aviva and the University of East Anglia underpinned a five-year research programme launched in 2016 by the IFoA's Actuarial Research Centre, bringing together actuarial approaches and health data. As reported in *Longevity Bulletin* 9, 'Big data in health', a multi-disciplinary group under Professor Elena Kulinskaya developed the MyLongevity app using anonymised EHR data from THIN to calculate life expectancy, taking into account socio-demographic profiles and health conditions (Kulinskaya and Gitsels, 2016). The research output of this group also included an investigation of the contribution that the use of statins had made to improvements in life expectancy, illustrating the potential for EHR datasets to support the development of future mortality improvement assumptions.

In 2021 a new actuarial working party was established under Niall Fennelly looking at future engagement with EHR datasets. This group has been working with underwriters and academics to better understand how electronic reports based on individual EHR reports could provide the data needed by insurers to underwrite and assess claims without adding to the burden of GPs with laborious paper forms (Fennelly, 2021). So actuarial engagement has been renewed, but the question is whether actuaries are once again playing catchup on the use and analysis of EHR datasets.

The impact of Covid-19 in boosting the use of electronic health records

The harrowing nature of the early months of the Covid-19 pandemic forced us all to re-examine our working principles. We needed to understand the epidemiology of Covid-19 quickly, and to support any actions that could improve treatment or prevent the spread of infections. Innovation was encouraged, while new paradigms cast aside once immoveable obstacles. The power of EHR datasets to identify those most at risk was quickly realised, and extraordinary efforts were made to connect existing data pools and free up access to researchers (Lynn, 2022).

An unprecedented effort involving the Bennett Institute for Applied Data Science at the University of Oxford, the EHR research group at the London School of Hygiene and Tropical Medicine, NHS England, and TPP, a global digital health company, brought the OPENSAFELY dataset into being in April 2020. This dataset contained primary care health records for 40% of the UK population. BY 7 May 2020 they had demonstrated the value of OPENSAFELY by publishing the world's largest study into factors associated with Covid-19 deaths.

Then, in a gesture that broke with the previous operating model, the collaborative group provided free and open access to the OPENSAFELY platform for other research groups to further our collective understanding of Covid. Every project was reviewed by NHS England to ensure that it supported relevant research and planning activities in response to the Covid-19 pandemic; so far 140 projects have been approved – summaries of each can be found at **opensafely.org/approved-projects**. A truly extraordinary effort and success story!

First steps with machine learning

Access to these EHR datasets is only part of the transformation. In the past, research projects would use generalised linear models and other regression techniques to determine the relationship between primary/secondary outcomes, target variables and pre-specified co-variates. An increasing proportion of research studies today use a variety of machine learning (ML) techniques to explore the richness of data provided by EHR datasets. These ML techniques are able to identify complex patterns in patient health information and end up producing much more detailed explanations of the underlying relationships (Siwicki, 2022).

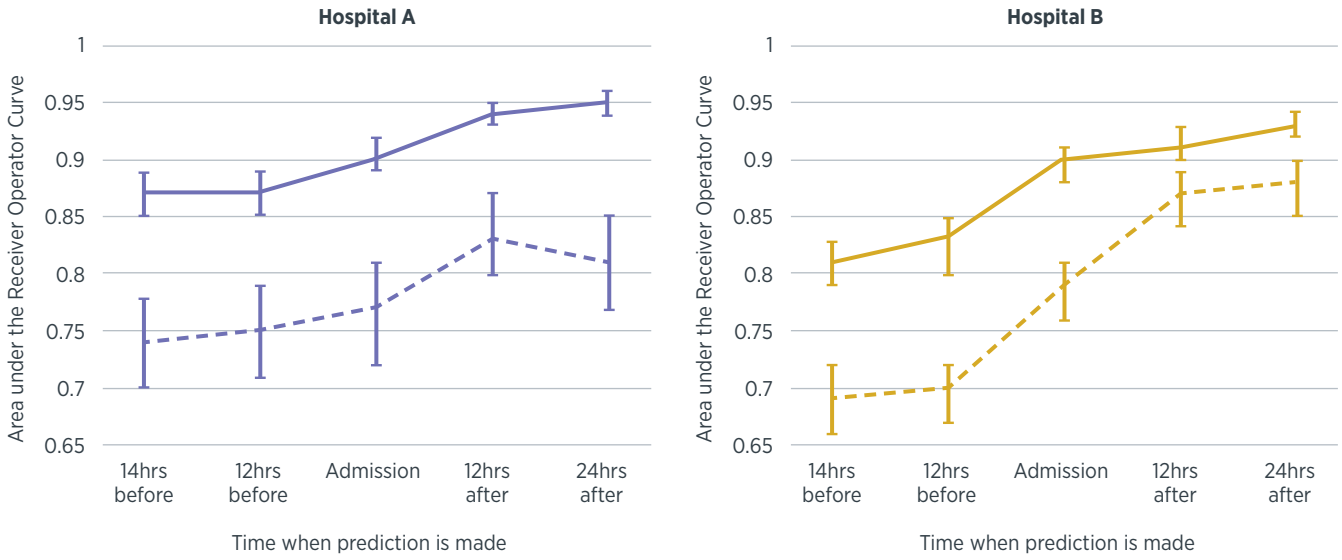
For example, ML models are able to rapidly construct time sequences of diagnostic codes, removing distortions caused by administrative processes. Combinations of diagnosis codes in a particular order may be more predictive than either the occurrence of diagnoses on their own or in a different order. This approach also has the advantage of identifying de novo disease markers that could prompt doctors to investigate further (Kent, 2020). However, care is needed in that improved techniques for identifying and quantifying a multitude of correlations do not provide a back-door to proving causation.

Further, we need to maintain some perspective over these impressive modelling capabilities, as ML patterns discovered in EHR datasets may reflect specific local features that it would be inappropriate to apply elsewhere. It is of vital importance that users of ML techniques require a combined understanding of the 'real-world situation' that the data is intended to represent, the circumstances of how the data was gathered and stored, and any assumptions and biases in the specific ML technique being applied. Such applications must involve rigorous training, testing and validation using different subsets of data, and any application of such techniques without due consideration poses a serious risk to the reliability of the outcome (Knevel and Liao, 2022).

By way of positive examples, Feng (2022) investigated the ability of a variety of ML techniques to predict Covid mortality and hospital admission risks. These ML techniques included gradient boost, random forest and AdaBoost. These supervised ML techniques were limited to routinely collected EHR data, and yet were able to achieve similar levels of predictive accuracy to regression analyses based on chest-imaging data that would only have been available after Covid-19 diagnosis.

Further, Rajkomar et al. (2018) used de-identified EHR data from two US medical centres with over 200,000 patients hospitalised for at least 24 hours, and were able to develop ML predictive models for inpatient mortality, unplanned readmission and discharge diagnoses using 46.8 billion data points. These ML models outperformed traditional predictive models over time, as illustrated in *Figure 1* on the next page.

Figure 1: *Comparison of ML and traditional predictive models on inpatient mortality for two US medical centres. (Source: Rajkomar, A., Oren, E., Chen, K., et al. (2018), Fig. 2. Licensed under* ***CC BY 4.0.****)*



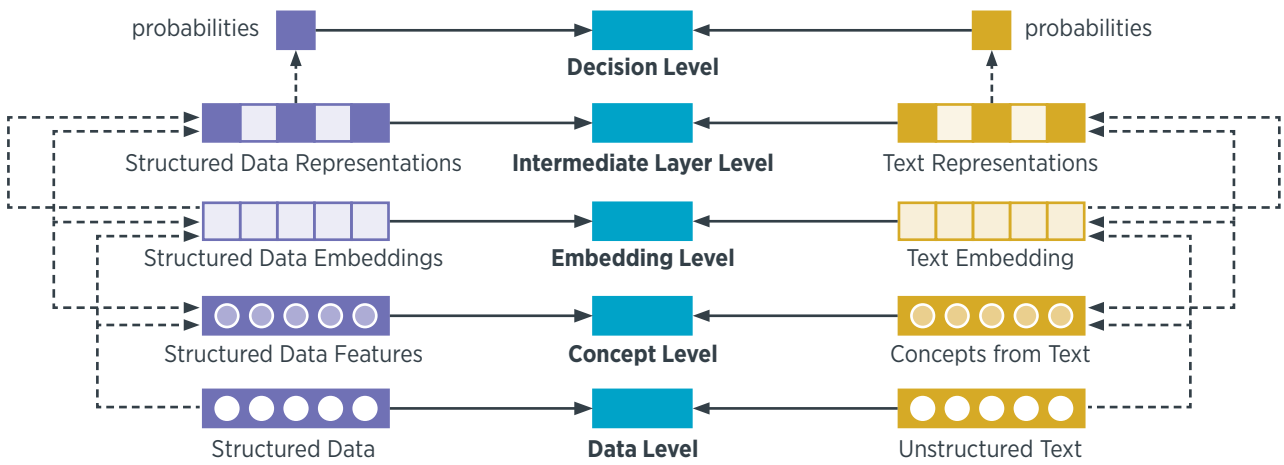
More and more research papers are investigating the possibility that the use of ML techniques, such as recurrent neural networks on EHR datasets, may provide doctors with the possibility of individual risk prediction and health trajectories. Ayala Solares et al. (2020) highlight the importance of focusing on EHR datasets that are sufficiently large and data rich, such as the CPRD, to be generalisable to other situations and on ML techniques that recognise the importance of time between events on the medical record.

Indeed, if we look at EHR datasets in more detail, we can see that reliance on traditional regression techniques constrains the data types that we can consider in developing predictive models. All EHR datasets consist of patient-level episodic data where diagnoses, tests and treatments are assigned to a date, and then combined first across all patients in a GP practice and then across all GP practices that contribute to that primary care software system provider.

As such, each episode consists of structured data, such as disease diagnosis codes, and unstructured data, such as free text, where GPs capture additional explanatory or qualifying text on patient's medical conditions and treatments. As there are no limits or guidelines as to what could be included in a free text field, traditional regression techniques cannot make use of the information provided (Liu et al., 2021).

As in other professions, GPs often adopt a shorthand language to allow information to be captured quickly and efficiently. As an initial first step, the THIN dataset identified commonly occurring typed phrases in the free-text fields, appending the relevant codes to each episode and hence converting some of the unstructured data to structured data. However, more advanced ML or deep learning (DL) techniques could use fusion strategies with more semantic layers to consider structured and unstructured data, or multimodal data, together and extract meaning across free-text fields rather than just individual words, as illustrated in *Figure 2*.

Figure 2: *Comparison of different interaction levels in analysing multimodal data. (Source: Liu, Z., Zhang, J., Hou, Y., et al. (2021), Fig. 2. Licensed under* ***CC BY 4.0.****)*



Wider clinical benefits from ML applications

However, it is not just researchers who are benefitting from the wider application of ML and DL techniques to EHR datasets. Increasing accuracy in the structure of medical coding means that even with training doctors waste valuable time identifying the correct code. Valuable time that could be spent interacting with the patient.

Efforts to develop ML models that help doctors to find necessary information in large EHR datasets require large and realistic databases of medical questions on which to train the models. As a layered approach in developing such databases, researchers at MIT have worked together with doctors to produce an initial database of 2,000 relevant questions, and then used this initial database to generate further questions using an ML model (Zewe, 2022). Evaluation of these further ML-generated questions indicated that they were of high-quality about 60% of the time.

Further efforts to reduce the amount of time spent by doctors in data entry have focused on advances in natural language processing. Digital assistants can now either use dictation to pre-populate key fields in medical records or extract meaningful content from background recordings during the patient consultation (Alldus, 2021).

Deeper engagement for insurers in machine learning

So what about the life insurance industry? What would be the benefits of deeper engagement in EHR datasets and the use of ML to develop predictive models? Would the benefits sufficiently outweigh the material costs that would be involved?

Let us first consider the current underwriting process. Underwriting forms and medical reports provide a detailed picture of the individual's current state of health. But the underwriting engine itself is likely the end result of combining different cohort studies into incidence and mortality associated with specific diseases. Combining these studies is a matter of expert judgement, as is considering what differences there may be between the participants of these cohort studies and real-world populations. These underwriting engines have clearly proved their worth, particularly if insurers evaluate their performance in broad buckets across different medical conditions with a similar level of impairment. The question is what more could investing in ML analysis of EHR datasets provide?

First of all, these large EHR datasets would provide actual matched experience to the underwritten lives in terms of medical conditions and treatments, rather than hypothecated experience. These EHR datasets are being continually updated, rather than relying on cohort studies that may be many years out of date. Further, these EHR datasets would provide direct information to underwriters as to how particular conditions are being treated, and how that compares to clinical guidance and previous treatment regimens. Moreover, access to free text and the use of ML techniques to extract semantic meaning from the free text would enable underwriters to compare prognosis for different severities of the same medical condition. Broad indicators of socio-economic status in the EHR dataset such as HealthACORN and MOSAIC could be enhanced by unfettered consideration of other proxies, and hence improve the relevance of EHR datasets to insured portfolios.

While not underestimating the costs involved, there could be a significant first-mover advantage in using real-world evidence to provide more precise estimation of the risks and price accordingly. Late adopters would run the risk that their aggregate pricing would be inappropriate and insufficient for the lives that they were insuring, and be late to notice and understand changes in prognosis for particular medical conditions.

Higher resolution in underwriting decisions may encourage actuaries to shift in developing future assumptions from top-down models of population mortality trends to bottom-up approaches that track the experience of limited numbers of statistically credible groups with similar medical impairments. For each group, EHR datasets would provide direct insights on potential sources of improvement or deterioration, either through considering differences between actual treatments or current clinical guidance, or changing understanding of the determinants of disease diagnosis. The direction of travel would be towards individual predictions of future improvements/deteriorations to complement more accurate underwriting assessments of current health.

The reality is that the wave of innovation that the Covid-19 pandemic unlocked has transformed our expectations of what is accessible and open to investigation. It has supercharged our investigation of the benefits of ML in understanding EHR datasets. The days of life insurance lagging behind general insurance when it comes to deep appreciation of data-rich environments may be coming to an end.

The UK Biobank, for example, contains a wide spectrum of health and mortality data on over 500,000 participants and is prepared to consider applications from both academia and industry if researchers can demonstrate how their research benefits public health and are prepared to publish. Bona fide researchers have the opportunity to work with different tiers of data with licensing costs between £3,000 and £9,000. Such opportunities need to be seized by the actuarial profession. It would be a great pity if actuaries were to be left behind by other data professionals just at the point that ML techniques start to unlock the potential of rich EHR datasets.

References

Alldus (2021). *Electronic Health Records (EHRs): how AI is improving clinician use*. **https://alldus.com/blog/articles/electronic-health-records-ehrs-how-ai-is-improving-clinician-use/** [Accessed 6 Feb 2023.]

Ayala Solares, J.R., Diletta Raimondi, F.E., Zhu, Y., et al. (2020). Deep learning for electronic health records: a comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, 101: 103337. **https://doi.org/10.1016/j.jbi.2019.103337**

Feng, A. (2022). A machine learning pipeline for accurate COVID-19 health outcome prediction using longitudinal Electronic Health Records. *AMIA Annual Symposium Proceedings,* 2021: 448-56. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8861740/** [Accessed 6 Feb 2023.]

Fennelly, N. (2021). The patient data pipeline. *The Actuary*, Oct.: 31-33. **https://www.theactuary.com/features/2021/10/04/patient-data-pipeline** [Accessed 6 Feb 2023.]

Kent, J. (2020). Machine learning tracks EHR Data to predict disease risk. *Health IT Analytics*, 19 June. **https://healthitanalytics.com/news/machine-learning-tracks-ehr-data-to-predict-disease-risk** [Accessed 6 Feb 2023.]

Knevel, R. and Liao, K.P. (2022). From real-world electronic health record data to real-world results using artificial intelligence. *Annals of the Rheumatic Diseases*, 82(3): 306-11. **https://doi.org/10.1136/ard-2022-222626**

Kulinskaya, E. and Gitsels, L. (2016). Use of big health and actuarial data for understanding longevity and morbidity risk. *Longevity Bulletin*, 9: 16-19. **https://actuaries.org.uk/media/xodfyuxb/longevity-bulletin-issue-9.pdf** [Accessed 17 July 2023.]

Liu, Z., Zhang, J., Hou, Y., et al. (2021). *Machine learning for multimodal Electronic Health Records-based research: challenges and perspectives*. **https://arxiv.org/abs/2111.04898** [Accessed 6 Feb 2023.]

Lynn, J. (2022). 2022 predictions for AI and machine learning in healthcare. *Healthcare IT Today*, 28 Jan. **https://www.healthcareittoday.com/2022/01/28/2022-predictions-for-ai-and-machine-learning-in-healthcare/** [Accessed 6 Feb 2023.]

Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1. **https://doi.org/10.1038/s41746-018-0029-1**

Siwicki, B. (2022). What's holding machine learning back in healthcare? *Healthcare IT News*, 12 Aug. **https://www.healthcareitnews.com/news/whats-holding-machine-learning-back-healthcare** [Accessed 6 Feb 2023.]

Zewe, A. (2022). Machine-learning models that can help doctors more efficiently find information in a patient's health record. *TechXplore*, 14 July. **https://techxplore.com/news/2022-07-machine-learning-doctors-efficiently-patient-health.html** [Accessed 6 Feb 2023.]

Dan Ryan

Dan Ryan is a medical demographer with three decades of experience in the insurance industry and is now Director of Demographic Risk at Just. He has led global multi-disciplinary teams at Willis Towers Watson and Swiss Re in diverse areas including forward-looking risk models, behavioural science, and how the combination of data science and digital ecosystems will transform how risk is assessed, managed and mitigated.

Dan has an MA in Medical Sciences from Cambridge University and an MBA from Heriot-Watt University. He is currently engaged in a DHealth at University of Bath that is examining the potential for better medication adherence to improve the management of hypertension through applying machine learning to electronic health records datasets such as the Clinical Practice Research Datalink.

The CMI's use of GLMs in the analysis of mortality and morbidity experience

The Continuous Mortality Investigation (CMI) performs regular analyses of UK mortality and morbidity data and makes use of a variety of techniques to support its outputs. In this article, we explore the CMI's use of generalised linear modelling (GLM).

Background

GLM is a flexible framework for analysing a variety of data types based on an extension of the linear regression modelling technique. Using GLM, we can model a relationship between a number of explanatory variables and the variable of interest, isolate the impact of each explanatory variable on the target variable and predict outcomes for that variable.

GLMs are used in many areas of actuarial work. The CMI often undertakes GLM analysis alongside its more conventional one-way analyses of mortality, using Actual/Expected (A/E) values to gain a more complete picture of the factors driving mortality rates.

GLM analyses allow a range of mortality and morbidity factors to be taken account of within a single model. They can be useful in understanding which factors are most significant and considering how they interact with each other. They can also highlight where one-way A/E values could be misleading and provide an alternative view of the true underlying experience.

The CMI takes a pragmatic approach to GLMs, in particular

- The results should be easily interpretable by the user of the output in which the analysis is published, so more complex models that include several interactions between explanatory variables are carefully considered to check whether they add sufficient insight to simpler models. For this reason, the CMI also often includes an offset term of expected claims/deaths, so that model coefficients can be shown as percentages centred around 100%, to aid interpretation.
- Data limitations can often constrain the robustness of some of the results. For example, dependence between variables that are assumed to be independent can occur where one data contributor only sells a particular type of business or through a particular sales channel. The CMI is therefore prudent in its interpretation and commentary of the results.

Two examples of GLMs used by the CMI include mortality investigations for pension annuities in payment and for term assurances, where the tool has added explanatory insight to more traditional one-way analyses.

Analysis of pension annuities in payment

The CMI Annuities Committee has used GLMs as part of their analysis, most recently in **CMI Working Paper 165**, as part of an analysis of mortality experience of individual annuities in payment. This analysis compared enhanced annuities, which are subject to an enhancement to the income paid to reflect health or lifestyle factors of the annuitant that are likely to lead to heavier mortality, with those receiving no specific enhancement ('standard' annuities). The analysis also considered the impact of the 2015 pension freedoms reforms on the individual pension annuity market.

GLM analysis was used to provide an additional perspective alongside traditional A/E analysis. The GLM model used for this analysis included the following variables:

- Annuity type (enhanced or standard)
- Commencement period (in this case, focusing on whether the annuity came into payment prior to pension freedoms being announced in March 2014, after they came into effect in April 2015, or in the transitional period between being announced and being introduced)
- Duration (length of time since the annuity came into payment)
- Annuity amount
- Office (the life insurance provider)
- Calendar year.

The GLM generally corroborated the one-way A/E analysis, by showing similar effects for several factors, but there were some differences of note. For example, the differential between experience of standard and enhanced annuities appeared greater in the GLM analysis than in the A/E analysis, and durational effects appeared to be stronger in the GLM analysis.

One interesting feature was that GLM analysis suggested that annuities commencing post-pension freedoms had experienced heavier mortality than annuities commencing pre-pension freedoms, while the one-way A/E analysis suggested the opposite. We conjecture that this may be due to the allowance for durational effects in the GLM analysis. This was an interesting analysis and one that we will consider repeating in future.

Analysis of term assurances

The CMI Assurances Committee has recently used GLMs as part of their work to produce the '16' Series term mortality and accelerated critical illness tables, published in **CMI Working Paper 150**.

GLMs were first used as a method to determine the drivers of mortality and morbidity to reflect in the tables, using a preliminary set of rates to calculate expected claims, which were included as an offset term. The analysis showed that:

- The factors that were reflected in the previous '08' Series tables – age, sex, smoker status and duration – were important factors that should be reflected in the graduated tables.
- It was reasonable to graduate the 'all offices' data – ie, it was not the case that any life insurance provider had rates with a substantially different age or durational shape to the preliminary rates.

More significantly, the analysis did not produce clear evidence that the shape of claim rates by age or by duration varied consistently across offices by other factors – ie, distribution channel, sum assured band, product type, joint / single life status and year of commencement. As a result, we decided not to vary the shape of the tables by any of these factors, as level adjustments to the tables should be adequate.

The Committee also used GLMs, with the proposed tables used to calculate expected claims, which were used as an offset term in the models. The results of these revised models were shown in the working paper to help users understand how experience varied by the factors that were not included in the rates. Results relating to individual life insurance providers' specific experience were collated in 'benchmarking documents' and shared with the data contributors to help them understand how their experience compared to the new tables.

Update on other items

The following provides an update on other recent CMI releases and upcoming work.

- The weekly mortality monitors, which are publicly available on the **CMI website**
- CMI_2022 was published alongside **Working Paper 177** on 22 June 2023
- "All offices" experience of term assurances in 2021 was published in **Working Paper 176**
- Analysis of long-term historical mortality improvements in **Working Paper 175**
- Proposed methods for the 'S4' Series pensioner mortality tables in **Working Paper 174**.

Institute and Faculty of Actuaries

Beijing

Room 512 · 5/F Block A · Landgentbldg Cente · No. 20 East Middle 3rd Ring Road
Chaoyang District · Beijing · 100022 · People's Republic of China
Tel: + 86 10 5878 3008

Edinburgh

Level 2 · Exchange Crescent · 7 Conference Square · Edinburgh · EH3 8RA
Tel: +44 (0) 131 240 1300

London (registered office)

1-3 Staple Inn Hall · High Holborn · London · WC1V 7QJ
Tel: +44 (0) 207 632 2100

Malaysia

Arcc Spaces · Level 30 · Vancouver suite · The Gardens North Tower
Lingkaran Syed Putra · 59200 Kuala Lumpur
Tel: +60 12 591 3032

Oxford

Belsyre Court · 1st Floor · 57 Woodstock Road · Oxford · OX2 6HJ
Tel: +44 (0) 207 632 2100

Singapore

Pacific Tech Centre · 1 Jln Kilang Timor · #06-01 · Singapore 159303
Tel: +65 8778 1784

www.actuaries.org.uk