BRIDGE TO TOMORROW
2023 INTERNATIONAL CONGRESS OF ACTUARIES
28 MAY – 1 JUNE 2023 · SYDNEY
Actuaries Institute Australia
IAA AAI International Actuarial Association Association Actuarielle Internationale

Drivers of Mortality- A Study using Artificial Intelligence and Machine Learning Techniques

Prepared by Iris Deng,
Simon Kong,
Lin Gu

Presented to the
2023 International Congress of Actuaries
28 May – 1 June 2023

© ***Iris Deng, Zurich Financial Services, Australia***
Simon Kong, Allianz Retire+, Australia
Lin Gu, RIKEN Center for Advanced Intelligence Project, Japan

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

1. Introduction

Mortality assumptions and life expectancy have important implications in pricing and valuation of life insurance products, particularly for products with periods of premium guarantees. Understanding drivers of mortality also plays an important role in product design, underwriting and risk management of life insurance products.

Advancements in the field of Artificial intelligence (AI) have led to the adoption of machine learning models in a wide domain of applications. While machine learning is known for its predictive capabilities, most of the models suffer from a relative lack of explainability and are based on correlation of parameters instead of causation. The causal reasoning algorithm is one of the latest advancements in AI and machine learning techniques that attempts to establish causal relationships and provide a basis for sophisticated human-like analysis of interventions and counterfactuals. Causal machine learning provides more precise and reliable estimates of causal effects.

This paper aims to propose a casual inference framework and to investigate the potential causal relationships between mortality and risk factors. Section 2 of this paper will briefly summarise the machine-learning literature. Section 3 of this paper will focus on data and methodology. Section 4 of this paper will apply a causal forest model to predict mortality rates and identify treatment variables which have a causal relationship with mortality. Section 5 outlines the application of causal reasoning and implications for life insurance and Section 6 concludes and discusses further research.

2. Literature Review

Victor et al. (2016) applied to learn the average treatment effect in an instrumental variables setting. It uses "double machine learning" method as it relies on estimating primary and auxiliary predictive models. The paper discusses the mathematical theory of double machine learning and introduces cross-fitting to prevent the loss of efficiency. It also introduces an empirical example of estimation of 401 (eligibility) on accumulation assets. Angusto et al. (2020) applied machine learning methodology to build a more realistic counterfactual scenario of mortality model and use it to derive estimates of local excess mortality during COVID-19 period. Qiao et al. (2021) used the structural causal model aided by machine learning methods to examine the potential causal relationships between COVD-19 severity and 10 environmental factors and results suggest that environmental factors are unlikely to cause noticeable aggravation of the COVID-19 pandemics. Fiona et al (2010) used machine learning technique to work out Energy Efficiency. This paper will mainly use double machine learning method and casual inference framework to work out the casual effect of treatment variables on the mortality.

3. Methodology and Data

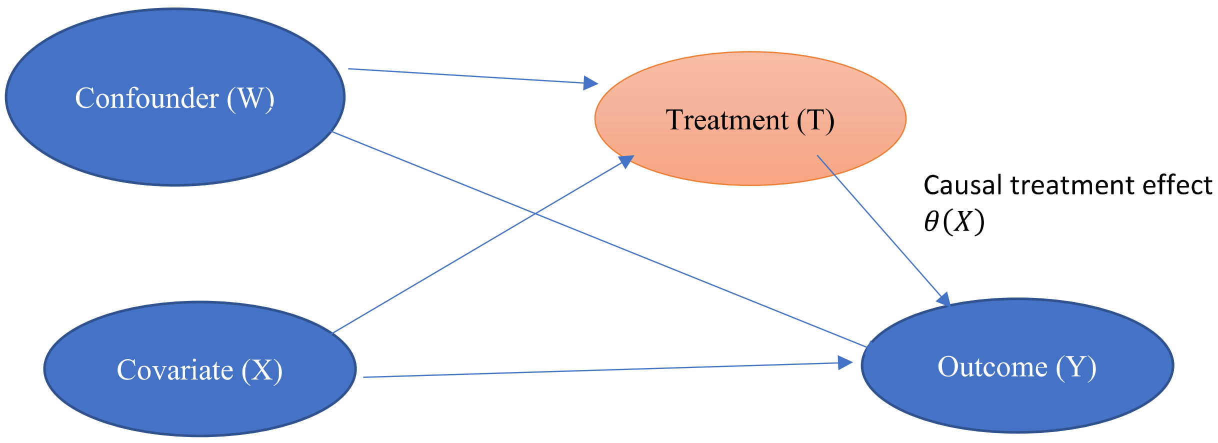
Causal Inference

Life insurance uses linear regression and other traditional statistical methods to assess the effect of risk factors such as gender, age, smoker status, BMI etc. These observed effects are used in rate setting and application of premium loading and discount.

Applying traditional statistical methods on life insurance data can give misleading effects if the confounding relationship between risk factors and other covariates is not eliminated. In a controlled experiment, confounding influences can be removed by randomisation of treatment and control groups such that any differences between the two groups can only reflect the causal treatment effect. However, this approach may not be practical for life insurance data. Without randomisation, applying traditional statistical methods have the potential of giving observed effects that are misleading and inequitable when used as the basis for assumption setting and premium differentiation.

Causal inference seeks to understand the underlying causal relationship between treatment and outcome, while allowing for the effects of other confounding variables on both the treatment and outcome. It identifies the relationship between variables by controlling for other factors that may be affecting the outcome.

Causal inference used in this paper follows the following structure:



Whereas:

- Causal impact is defined as the magnitude by which the outcome (Y) is changed by a unit-level interventional change in treatment (T).
- Covariates (X) and confounders (W) are factors that simultaneously influence the likelihood of the treatment (T) and the outcome (Y).
- The causal treatment effect of T on the outcome (Y) is defined as a function of X i.e. $\theta(X)$.

In the context of mortality modelling, mortality rate is the outcome; the treatment is the selected risk factor under investigation; gender, age and other common risk factors are assigned to covariates and confounders that influence both the treatment and mortality rate.

Double Machine Learning (DML)

Double Machine Learning (DML) uses a two-stage approach for estimating the treatment effect on the outcome. The first stage of DML estimates two predictive relationships:

1. Predicting the outcome (Y) from confounders (X and W)
2. Predicting the treatment (T) from confounders (X and W)

The second stage of DML combines these two predictive models in a final stage estimation to create a model of the heterogenous treatment effect.

Formally, the DML model takes the following structural equations:

$$Y = \theta(X) * T + q(X, W) \; + \; \epsilon \qquad \mathbb{E}[\epsilon | X, W] = 0$$

$$T = f(X, W) + \eta \qquad \mathbb{E}[\eta | X, W] = 0$$

$$\mathbb{E}[\eta * \epsilon | X, W{=}0]$$

The goal is to estimate the Conditional Average Treatment Effect (CATE) $\theta(X)$. The equations can be rewritten as:

$$Y - \mathbb{E}[Y | X, W] = \theta(X) \cdot (T - \mathbb{E}[T | X, W]) + \epsilon$$

The first stage of DML estimates the conditional expectation functions:

$$q(X, W) = \mathbb{E}[Y | X, W]$$

$$f(X, W) = \mathbb{E}[T | X, W]$$

This leads to the residual equations:

$$\tilde{Y} = Y - q(X, W)$$

$$\tilde{T} = T - f(X, W) = \eta$$

And

$$\tilde{Y} = \theta(X) \cdot \tilde{T} + \epsilon$$

The second stage model performs the final regression problem:

$$\tilde{\theta} = \operatorname{argmin}_{\theta} \mathbb{E}\left[\left(\tilde{Y} - \theta(X) \cdot \tilde{T}\right)^2\right]$$

Causal Forest DML

Casual forest is a machine learning method from generalized random forests (Athey et al., 2019). A causal forest is built from causal tress, where the causal tress can learn a low-dimensional representation of treatment effect heterogeneity. It aims to find out the casual effect between the treatment variables and predicted variables. It is an extension of random forest. In random forest, data is repeatedly split in order to minimize prediction error of an outcome variable. Causal forests are built to maximize the difference across splits in the relationship between the treatment and the outcome.

In this paper, we implement Causal Forest as the second stage model of DML. The estimation of $\theta(X)$ is based on solving a local residual on residual moment condition:

$$\tilde{\theta}(x) = \operatorname{argmin}_{\theta} \sum_{i=1}^{n} K_x(X_i) \cdot (Y_i - \tilde{q}(X_i, W_i) - \theta \cdot (T_i - \tilde{f}(X_i, W_i)))^2$$

Selection of First Stage Model

Two first stage models are considered in this paper:

Least Absolute Shrinkage and Selection Operator (LASSO) Regression

Linear regression is a standard modelling technique commonly used in the insurance industry. LASSO regression is a type of linear regression method that performs both variable selection and regularisation in order to enhance the prediction accuracy and interpretability of the resulting statistical model. It is selected over other linear regression methods to reduce the risk of overfitting.

Random forest

Random forest is a commonly-used ensemble machine learning method that operates by constructing a multitude of decision trees at training time and combining the output from those decision trees to reach a single result. For regression tasks the mean prediction result is returned.

Data

Population data by Statistical Area Level 2 (SA2) 2016 and 2011 is sourced from the Australian Bureau of Statistics (ABS) and the Australian Institute of Health and Welfare (AIHW). Specifically, the following data are collected for modelling purpose:

- Number of deaths and death rates
- Various risk factors which might be related to mortality rates, including age, gender, prevalence for different chronic disease, social economic index, exercise, education level and employment etc.

SA2 2016 is defined by the Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016. It consists of 2310 regions in Australia in 2016 and 2214 regions in 2011 without gap or overlap respectively. Each SA2 generally has a population range of 3,000 to 25,000 persons. Individual SA2's with small population size are removed. When SA2 data is not available, data by Population Health Area (PHA) and Statistical Area Level 3 (SA3) are mapped to SA2.

Model Parameterisation

Hyperparameter tuning is performed for each of the first stage models and then again for causal forest. The final model is selected based on the overall mean squared error (MSE) of DML model.

Backward elimination with ordinary linear regression is then applied to the data to reduce the number of variables for mortality modelling from 207 to 43. The table below shows the grouping of these variables into covariates X's, confounders W's and treatments T's. Specifically, age, gender and health risk factors and disease prevalence/incidence are covariates (X's) in the model, treatments (T's) are selected variables that are being assessed in the paper for causal effect; Other influencing factors are assigned to confounders (W's). Note that Y is mortality rates.

W	**X**
Income units receiving Commonwealth Rent Assistance	age 30 below percentage
fully immunised children percentage	age 30 to 49 percentage
Asian percentage	age 50 to 69 percentage
fully engaged percentage	age 70 to 79 percentage
Bachelor Degree and Above percentage	age 80 to 89 percentage
Non-indigenous percentage	male percentage

Married percentage	chronic kidney disease prevalence rate per 100
Public transport to work percentage	diabetes prevalence rate per 100
Vehicle to work percentage	heart disease prevalence rate per 100
out of pocket cost per imaging service	lung condition prevalence rate per 100
out of pocket cost per GP	mental health prevalence rate per 100
people who did not claim a GP attendance percentage	stroke prevalence rate per 100
woman who are pregnant smoker percentage	obesity percentage
speak English well and very well percentage	current smoker percentage
white collar percentage	adequate fruit intake percentage
Light blue percentage	colorectal cancer incidence rate per 1000
Internet accessed from dwelling percentage	melanoma of skin incidence rate per 1000
Owned outright and owned with a mortgage percentage	lung cancer incidence rate per 1000
Specialist attendances per person	all cancer incidence rate per 1000
Medicare benefits expenditure per person	all other cancers incidence rate per 1000
Total out-of-pocket cost per patient for non-hospital Medicare services	insufficient physical activity prevalence per 100
unpaid domestic work less than 5 hours per week percentage	

To estimate the effect of each treatment, a new model fit is performed which includes hyperparameter tuning of first stage and causal forest models based on the selected treatment. This may lead to different first stage models being selected for modelling treatments (and mortality).

4. Results and Implication

This section provides results of the model fitting. LASSO is used as the first stage model for Causal Forest DML. The table below sets out the CATE and p-value for each treatment variable analysed in this paper. MSE and R-squared values of the fitted models (found in the Appendix) indicates that Causal Forest DML generally provides a slightly better fit than the standalone LASSO model. In addition to the small improvement in MSE and R-squared, Causal Forest enables the analysis of causal treatment effect whereas traditional methods tend to focus on correlation only.

Treatment Variable	CATE [1]	P-value
insufficient physical activity	3.80	<0.05
mental illness	11.95	<0.1
adequate fruit intake	-0.028	<0.1
smoker	4.45	<0.15
diabetes	0.067	<0.15
kidney disease	0.174	<0.2
obesity	0.037	<0.2
heart disease	0.03	>0.2
lung disease	4.2	>0.2

[1] *CATE of 1 represents an average increase in mortality of 1 per thousand from the treatment.*

The results indicate statistically significant causal effect from insufficient physical activity, fruit intake and mental illness. Smoker, diabetes, kidney disease and obesity have p-values of 0.2 or less. Other factors such as heart and lung diseases are found to be statistically insignificant.

Key insights from the results are discussed in detail below.

Insufficient Physical Activities vs Obesity

Insufficient physical activity[1] is defined here as:

- for adults aged 18–64: not completing 150 minutes of moderate to vigorous activity (where time spent on vigorous activity is multiplied by 2) across 5 or more days a week

---

[1] AIHW 2022, https://www.aihw.gov.au/reports/australias-health/insufficient-physical-activity

- for adults aged 65 and over: not completing 30 minutes or more of physical activity on at least 5 days each week.

Obesity[2] in this paper is defined as having Body Mass Index (BMI) of greater than 30.

BMI is a commonly used rating factor in life insurance underwriting as it is easily measurable. A person with obesity or high level of BMI is usually associated with poorer mortality expectation and is more likely to attract a premium loading than policyholders with lower BMI.

However, the causal analysis suggest that obesity is not a particular good measure of mortality with CATE of 0.037 and p-value of 0.2.

The causal analysis also shows that insufficient physical activity is a more prominent risk factor for mortality with CATE of 3.8 and p-value <0.05. This is supported by numerous medical research papers that concluded that:

- higher physical activity was associated with lower risk of mortality irrespective of weight status. Compared with obesity-low physical activity, there was no survival benefit of being normal weight if physical activity levels were low. (Jakob et al 2022)
- high physical activity was also associated with lower risk of all-cause mortality in nearly all strata of adiposity, irrespective of the adiposity measure used, with the exceptions of BMI greater than 35 kg/m2 and low-Body Fat percentage. (Miguel et al 2020)
- focusing on improving cardiorespiratory fitness, which is largely driven by physical activity and exercise, and reducing visceral/ectopic adiposity, may be more important, as these are the key drivers of cardiometabolic diseases and adverse outcomes in patients with overweight and obesity. (Carl et al 2022)

For insufficient physical activities, the causal effect is estimated to be 3.8 per 1000 using causal forest, compared to 2.6 per mile increase using linear regression methods.

Mental Illness

Mental illness has been one of the biggest drivers of recent poor morbidity experience in Australia. This paper shows that it is also a prominent risk factor for mortality with CATE of 11.5 per thousand and p-value of 0.085. The result suggests that mentally ill people, on average, have mortality rates that are approximately two times higher than people without mental illness. According to the Royal Australian and New Zealand College of Psychiatrists, the higher mortality from mental illness may be attributed to factors such as greater exposure to other risk factors; reduced access to healthcare due to financial barriers and discrimination among healthcare providers;

---

[2] AIHW 2022, https://www.aihw.gov.au/reports/australias-health/overweight-and-obesity

systemic issues in health-care delivery; adverse effects of medications; and lack of capability among health professionals to deal with complex comorbidities.

The magnitude of the causal effect is supported by the following research papers:

- A meta-analysis of studies worldwide (Walker et al. 2015) estimated that people with mental illnesses had a mortality rate 2.2 times that of people without, and an average of 10 years of potential lost life. There is also evidence that this gap is increasing (Firth et al. 2019).
- Research from Western Australia found that the gap in life expectancy increased between 1985 and 2005. Notably, like other research, only a small portion (14%) of the gap in life expectancy in this study was attributed to suicide. Most (80%) of the gap was attributed to physical health comorbidities, such as cardiovascular disease, respiratory disease and certain cancers (Lawrence et al. 2013).

For mental illness, the causal effect is estimated to be 11.5 per 1000 using causal forest, compared to 10.8 per mile increase using linear regression methods.

Adequate Fruit Intake

Adequate fruit intake is defined as the proportion of people eating sufficient serves of fruit every day. Negative relationship between fruit intake and mortality is widely recognised by health professionals and supported by numerous medical research papers.

This is confirmed by the causal analysis in this paper which indicates a statistically significant causal effect (p-value =0.067); having adequate fruit intake is estimated to reduce mortality (albeit only by 0.028 per thousand). The small causal effect of 0.028 per thousand is likely the result of limited sample size and lack of geographical variation in the level of fruit intake across Australia.

Smoker

Smoker is a rating factor in retail life insurance as smokers have higher death rates compared to non-smokers. Our results indicate that smoking, on average, increases mortality by 4.5 per thousand. However, the p-value is 13.6%. This might be again due to limited sample size and lack of geographical variation in the proportion of smokers across Australia. A larger dataset is likely to lead to a lower p-value for this risk factor. For smoker, the causal effect is estimated to be 4.5 per 1000 using causal forest, compared to 0.57 per mile increase using linear regression methods.

5. Application

This paper explored the application of the causal inference framework and use of causal forest machine learning techniques to investigate the causal relationship between risk factors and mortality. While traditional statistical methods used in life insurance may give reasonable predictions of the mortality experience, those methods focus on correlations instead of underlying causal relationship. They also lack the ability to control for confounders that can bias the estimate of the treatment effect. Without proper expert opinion, this could lead to inappropriate risk factors (and sequencing of those risk factors) being used for rate setting purpose.

Causal forest discussed in this paper has the potential of addressing this risk and improving equity among policyholders by examining the underlying causal effect and selecting the appropriate risk factors for premium discrimination. Compared to traditional regressions methods, causal forest can give causal effects that are substantially different. As demonstrated in the Result section, the causal forest analysis (based on the population data) gave causal effect of 3.8 per thousand for insufficient physical activity compared to 2.6 per thousand from linear regression.

Analysis of underlying causal effects enables insurers to make more informed decisions on the cost and benefit trade-off of potential interventions to improve the mortality outcome. For instance, an insurer can look at the potential saving in mortality cost relative to the cost of interventions which can include discounts on gym membership, reward program based on activity tracking and promotion of healthy lifestyle habits. It is noted that there are currently similar programs offered in the market by AIA (i.e. Vitality) and Zurich (i.e. LiveWell) that promote healthy lifestyles. Application of the causal inference framework and causal forest machine learning techniques can improve the cost effectiveness of such program by refining the reward programs to target the underlying causal relationships.

6. Conclusion and Further Research

This paper demonstrated the feasibility of applying the causal inference framework for life insurance. Greater explanatory power delivered by causal models can improve understanding of real causes of claims experience and deliver fairer pricing outcomes for customers. It can also inform product designs and improve cost effectiveness of interventions (e.g. reward programs) to improve claims outcome.

However, it should be noted that the analysis in this paper was based entirely on Australian general population data that were freely accessible. No proprietary data was used. Analysis results presented in this paper may not necessarily reflect the proprietary claims experience based on the insured population. The analysis in the paper was also limited by the size and quality of available data. Further enhancement and research can include:

- Consideration of additional risk factors.
- Performing the analysis on the insured population data.
- Segmentation of data for analysis by age group and gender.
- Use of other first stage models and alternative causal machine learning models

Appendix

The table below compares the MSE and R-Squared for predicting mortality rates for test data by using random forest, LASSO (as a standalone model) and Causal Forest DML (with LASSO being the first stage model) using insufficient physical activity as a treatment variable. We also used average value as a predict for test data for comparison purpose.

	simple average	**Random forest**	**LASSO**	**Causal Forest**
Mean Square Error (MSE)	12.6	2.84	2.87	2.84
R-squared	NA	0.78	0.78	0.84

It shows that Causal Forest DML gives a slightly better fit than the standalone LASSO model. The key benefit of Causal Forest is its ability to analyse causal treatment effect more accurately than traditional methods that tend to focus on correlation only. Under the causal forest model, insufficient physical activities is estimated to increase mortality rate on average by 3.8 per thousand, compared to 3.0 per thousand increase from LASSO model.

Reference

Susan A, Julie T & Stefan W (2019), Generalized Random Forests, Cornell University, *ArXiv e-prints*, <https://arxiv.org/abs/1610.01271>

Victor C, Denis C, Mert D, Esther D, Christan H, Whitney N and James R (2016), Double Machine Learning for Treatment and Causal Parameters. Cornell University, ArXiv e-prints

Cerqua, A, Di Stefano, R, Letta, M. et al. (2021) Local mortality estimates during the COVID-19 pandemic in Italy. *J Popul Econ* 34, 1189–1217

Qiao K, Xing S, Xiaying X, Bing C, Yuanzhu C, Xudong Y, and Baiyu Z (2021), Machine Learning-Aided Causal Inference Framework for Environmental Data Analysis: A COVID-19 Case Study, Environmental Science & Technology 2021 55 (19), 13400-13410

Fiona B, Christopher K, David R, Mar R, Catherine W (2020), Machine Learning from Schools about Energy Efficiency, The Association of Environmental and Resource Economists

Walker ER, McGee RE, Druss BG (2015). Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry

Jiska CM and Rotem P (2011), Is there a reversal in the effect of obesity on mortality in old age? Journal of Aging Research

Tarp J, Fagerland MW, Dalene KE (2022), et al, Device-measured physical activity, adiposity and mortality: a harmonised meta-analysis of eight prospective cohort studies British Journal of Sports Medicine; 56:725-732.

Miguel A, Sanchez L, Ding D (2021), physical activity and mortality across levels of adiposity, Mayo clinic proceeding

Firth J, Siddiqi N, Koyanagi A, Siskind D, Rosenbaum S, Galletly C et al. (2019), The Lancet Psychiatry Commission: a blueprint for protecting physical health in people with mental illness, *The Lancet Psychiatry*, 6(8):675–712, doi:10.1016/S2215-0366(19)30132-4.

Lawrence D, Hancock K and Kisely S (2013), The gap in life expectancy from preventable physical illness in psychiatric patients in Western Australia: retrospective analysis of population based registers- external site opens in new window, *British Medical Journal,* 346:f2539, doi:10.1136/bmj.f2539