Predicting mortality from 57 economic, behavioural, social, and psychological factors

Critical Review

Student: 2938740
Group 14

Predicting mortality from 57 economic, behavioural, social, and psychological factors

Article Selected:

Eli Puterman et al. Predicting mortality from 57 economic, behavioral [sic], social, and psychological factors, *PNAS July 14, 2020 117 (28) 16273-16282*; First published 22/06/2020. Link: https://doi.org/10.1073/pnas.1918455117

Part 1 Summary:

Introduction

Humans are a diverse species, and our individuality is often expressed as a celebration of uniqueness when compared against our peers. However, we all share two significant events in our lives - the day of our birth, and the day of our death. The journey we take as individuals between these events is also a celebration of our perceived uniqueness but our expiration, barring accident or foul play, is usually drawn from a common pot of maladies with heart attack, stroke and cancer being among the most common outcomes.

These biological factors are well researched and understood in terms of predicting mortality from a clinical perspective. Puterman et al. [1] seek to bridge the gap between clinical knowledge and policy decisions by examining a range of multidisciplinary factors in an attempt to provide future scientists and government policy makers with additional armaments in the quest to overturn a three decade long stagnation of US life expectancy figures. To this day, the US remains the only OECD developed country that does not provide universal health care coverage [2].

The study is a predictive analysis of 57 mortality factors across a transdisciplinary base of economic, socio-behavioural, and psychological factors using an advanced regression model and machine learning algorithms. Previous studies have tended to include biological factors when performing cross-domain research which only serve to obscure any insights that can be gained by focussing solely on the forementioned non-biological domains in relation to predicting the long-term factors which most influence mortality.

The goals of this study are to:

- Identify non-biological risk factors affecting mortality
- Identify which risk factors (if any) require prioritisation
- Generate novel research across multidisciplinary domains
- Influence government health policy with holistic thinking
- Guide future research based on the strength of the results
- Encourage development of lifespan models for future studies

Data Acquisition and Preparation

The primary data was sourced from open data provided by The University of Michigan Health and Retirement Study (HRS) [3]. The HRS is a nationwide longitudinal study based on a representative sample of 20,000 Americans and is supported by the National Institute on Ageing (NIA) and the Social Security Administration (SSA). The study has been collecting data via in-depth interviews on a bi-annual basis since its inception in 1992. The HRS website contains an enormous amount of pre-processed data and is classed as a data provider service for scientists and researchers.

The dataset for this study contained a cohort of 13,611 adults ranging from 52 to 104 years of age (mean 69.3 years old). The dataset for this study was collected between 1992 and 2008 and tracked 6 years of follow-up data per adult, or interviews with family where death occurred before 6 years. The primary dataset, whilst representative, does contain some racial (77% white) and gender (59% female) bias.

The study tested its regression model against a second, independent set of comparison data. This data was sourced from the Midlife in the United States (MIDUS) study [4], also supported by NIA grants. MIDUS was deemed to offer the most comprehensive set of variables in the US that best matched the social and behavioural factors considered in the primary data with 39 of the 57 HRS variables being present in the MIDUS data across all 6 transdisciplinary domains. There is no information about the size of the MIDUS dataset.

The data is split across six domains (childhood adversity, and 5 adulthood domains - socioeconomic, health behaviours, social connections, psychological and adverse experiences). The 57 predictors cover a range of negative experiences within the listed domains including smoking, alcohol abuse, state of mind, satisfaction, exposure to discrimination, relationships, education, income, etc.

Data Modelling and/or Analysis

The study used a detailed combination of established methods combined with cutting edge techniques to provide a rigorous examination of the data:

- Cox regression models
- Lasso regression
- Random forest prediction

The data was weighted and standardised where required and different techniques applied for continuous (mean and standard deviations) versus categorical (totals and percentages) variables. The Cox regression model produced hazard ratio and confidence intervals for each predictor, using Bonferroni multiple comparison corrections to improve the accuracy. A sensitivity analysis was also carried out to determine any skewing of the data from those who died within 2 years of the data sample end date (2008). Next, each domain was analysed in isolation to determine how much of the sample variance it explained. The top 3 dimensions that explained the most variances were chosen from each domain via Principal Component Analysis (PCA) and these informed another multivariate cox regression comparing individual domains, and as a combined set to determine any interaction effect. The model was applied to the MIDUS data to analyse/confirm the validity of the results.

As this study deals with many potential predictors, a lasso regression was performed to find a model that provided the best parsimonious results. This type of regression was deemed suitable given the lack of high correlation predictors that can lead to multicollinearity whilst penalising instead of

dropping variables (as per forwards/backwards model fitting techniques). Cross-validation was used to help select the final model.

Finally, a machine learning algorithm called random forest was employed to predict survival rates. The data was split 2/3 training and 1/3 test with bootstrapping utilised to improve the robustness of the results. The root nodes were split using the strongest predictor from each prediction. Final classification was made using the majority votes concept based on using all predictors within the model.

Results

A summary of the salient results with accompanying charts is presented in the main document along with a supplementary appendix document containing a very detailed set of results and breakdown of the data structure. The results in the main document are graphically represented via 1 table and 3 charts. They are consistent in being easily interpretable and show the predictors with the largest hazard ratios quite clearly. The results are taken from the Cox regression and the lasso regression models with none shown from the random forest. The data presented consists of the hazard ratio and confidence intervals. The captioning of the tables and charts clearly described exactly what you are looking at:

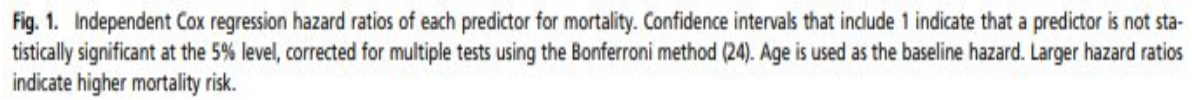
Fig. 1. Independent Cox regression hazard ratios of each predictor for mortality. Confidence intervals that include 1 indicate that a predictor is not statistically significant at the 5% level, corrected for multiple tests using the Bonferroni method (24). Age is used as the baseline hazard. Larger hazard ratios indicate higher mortality risk.

Figure 1: Example of descriptive captioning of research paper charts

Figure 3 (in paper) represents the biggest takeaway in how the results from the HRS and MIDUS studies overlap and are presented by domain, allowing the reader to instantly compare the validity of results and identify if any domain shows multiple predictors of concern as with adulthood health behaviours for example.

Each model selected predictors according to its own strengths but commonality was observed in the results. At a domain level, it was found that each domain added substantially to the explainable variance within the model. It was also noted that whilst some predictors appeared to be significant when viewed in isolation, the strength of their predictive abilities diminished when viewed as a collective of predictors with variance in this effect observed between models, for example, whilst Cox regression highlighted childhood psychosocial issues as important, the lasso regression ranked it higher in its model.

The study found that the results were not impacted by gender, race, or educational achievement. Age was found to be a factor in determining associations between predictors with differences seen at the above and below 75 years of age cut-off.

Conclusions, Actions or Decision Making

The results were largely consistent and sometimes matching across the regressions and datasets with the top predictors of mortality being smoking (current & historical), history of divorce, financial/employment difficulties, and alcohol abuse. Interestingly, feelings of hopelessness and history of renting appeared elevated which are particularly relevant in today's political landscape with increasing stress, access to property ownership and general unaffordability being rife across the western world.

The main conclusion is that a multitude of factors across several domains (behaviours, financial wellbeing, social and psychological) were significant influencers of mortality whichever statistical model was used, however that is not to say that the same predictors were identified by all models to the same extent. In lasso regression, some results ran counter to existing literature on the subject and the authors suggest that there may be an unknown set of factors influencing the data at a generational level, citing evidence of traumatic events such as the great depression and world war 2 and how they would present different obstacles and effects for each generation to overcome both physically and mentally.

No actions were presented in the discussion however they did re-affirm their goals of producing evidence for conducting future studies in new ways that embrace a more holistic lifespan approach for studies of the human condition. Liu et al [5] are quoted as finding that health and genetic markers may account for up to one third of the variation in mortality prediction data, but these markers were not included in this study. They also suggest including social trajectory theory in future research to account for contributions from childhood development impacting levels of success as an adult, as well as a contribution from macro level influencers such as systemic racism and local, provincial and national government policy.

References

[1] Eli Puterman et al. Predicting mortality from 57 economic, behavioral [sic], social, and psychological factors. *Proceedings of the National Academy of Sciences* Jul 2020, 117 (28) 16273-16282; DOI: 10.1073/pnas.1918455117

[2] Universal Health Care in Different Countries, Pros and Cons of Each - Why America Is the Only Rich Country Without Universal Health Care. Link: https://www.thebalance.com/universal-health-care-4156211

[3] Data source provider. Health and Retirement Study. Link: https://hrs.isr.umich.edu/

[4] Data source provider. Midlife in the United States study.
Link: https://www.nia.nih.gov/research/resource/midlife-united-states

[5] Z. Liu et al. Associations of genetics, behaviours, and life course circumstances with a novel aging and healthspan measure: Evidence from the Health and Retirement Study. *Plos Medicine*.
Link: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002827

Part 2 Critical Assessment:

Strengths

The main strength of this study is the rigorous application of data science methods and interrogation of the data. Steven Skiena [6] states in the introductory pages of the coursebook that one of the most fundamental principles is to "value doing the simple things right". The study starts by not only finding a reputable dataset of high quality but finds a second, high quality dataset against which to compare the results on a like-for-like basis, with both datasets taken from national studies and supported by major Government funding. Supplementary documents are provided to explain the data transformation and code based used in producing results.

The initial analysis uses an industry standard regression method (Cox) for time-based survival analysis before attempting to improve on the methodology through introducing a new regression method (lasso) and extensive machine learning concepts (random forest, boot-strapping and cross validation of results).

The study is very transparent in acknowledging its limitations and how other studies have generated results that are particularly relevant to the over-arching theme of inspiring future research to develop more complete lifespan models for assessing mortality.

The study generates some very clear results that are not entirely unsurprising by themselves (i.e. smoking is a major health hazard) but the real strength lies in the grouping of factors into specific domains such as adult behaviour, psychological issues, etc. It is this presentation of domain that will allow the greatest leveraging of this study not only in future research but also through applied policy lobbying.

Originality and Novelty

Whilst the authors make noise about introducing a new way to analyse mortality predictors and of aspirational aims to guide future research and policy decisions through a more holistic approach, the theory is not new. In 1977 George Engel [7] published a theory that called for a new medical model to understand a person's health in a more holistic way. The "biopsychosocial model" [8] outlined how the marriage of biology, psychology and sociology could be leveraged to explain the underlying factors that cause or exacerbate an individual's health conditions, rather than relying purely on biological pathology. To highlight progress in this field it has recently been accepted that fibromyalgia may be triggered by excessive stress [9] (emotional or physical) instead of some undiscovered biological fault. To show the power of this model, the acknowledgement that it can be caused by non-biological factors is simply a great comfort for sufferers.

Where the study does show novelty is in the fact that its very existence serves to rebuke one of the key criticisms of the biopsychosocial model – that it is not based in hard science [8]. Figure 2 [10] shows the core triumvirate of bio-psycho-social domains and it takes little imagination to see how an unbalancing of one domain could quickly affect not only the other domains, but also create collateral damage to the people closest to the individual, leading to a compounding of the original problem and possibly a vicious circle of disease, stress, and unhealthy behaviour, all impacting mortality.

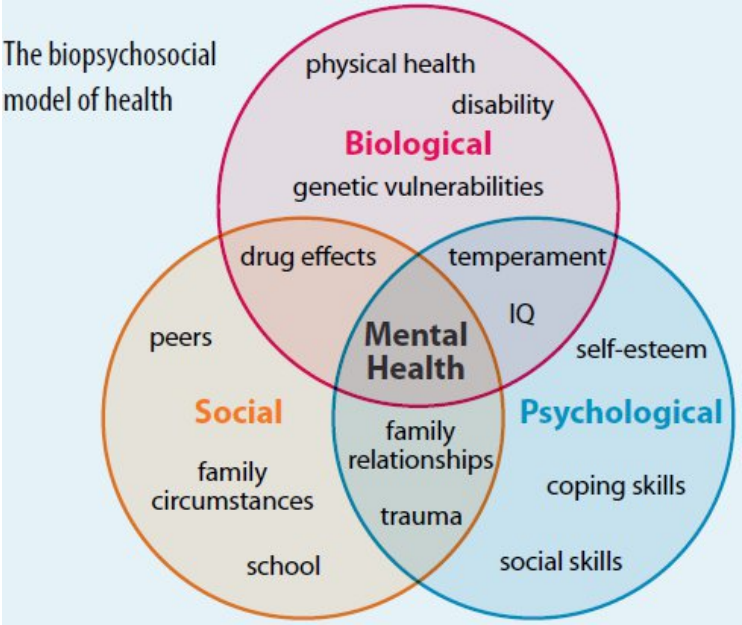

Figure 2: Interactions within the biopsychosocial model

Concerns and Limitations

The study highlights some concerns and limitations and discusses accordingly:

- Regarding the possibility of recall issues relating to childhood trauma they refer to a similar study by Hardt and colleagues' which states that evidence shows this is not an issue when the terms of the trauma are clearly defined. Hardt's position was supported by other nationally representative studies.
- The study presents a set of results which does not allow for a causal interpretation. A mortality factor, or set of factors within a domain, may have a significant effect on predicting mortality but the study does not propose a solution or reason as to why.
- The study admits that the adverse events allocated to some domains were not fully developed or were missing measures (food insecurity, domestic abuse) used in other studies. They recognise that differing study timescales may require different factors.

The ‘Methods’ section consists of 3 small paragraphs which feel like an afterthought with the actual methodology described in the ‘Results’ and ‘Statistical Analysis’ sections. Whilst well written they do blur the boundaries of what was expected, with the ‘Discussion’ section sandwiched in between them.

My main concern about this study is that it does not propose a serious set of follow up actions. The authors freely admit that study results are intended to serve as a tool for future hypothesis generation.

Data science guides [11] recommend that summary statistics are presented to form an expectation of the data however this was notably absent in the final report.

Potential Societal Impact

The biopsychosocial model provides the basis for an all-encompassing modern health framework and studies such as this can only help to provide the concrete evidence that modern medicine should be about more than simply dispensing pills. This approach is not without its parallels if we consider the approach taken in traditional Chinese medicine that the mind and body are interconnected. Western science is only starting to take this concept seriously and in 2018, neuroscientists were able to track a thought as it moved through the body of an epileptic patient undergoing surgery [12].

It stands to reason that increasing research into this area will produce a lot more evidence and insight into how our psychological conditioning and behaviours play into affecting not only our mortality, but also our general health. If we can identify the behaviours or adverse experiences that increase mortality and sickness, then we can instigate programs and policies that attempt to mitigate or identify at-risk individuals before the problem reaches a chronic or terminal outcome. Government policy (local and national) could then be structured in a way that relieves pressure on the individual and creates a more humane, cohesive, and less stressful society.

As a lifespan model there is an opportunity to remediate malignant attitudes before they take root and create future issues. Multiple studies have proven the existence of intergenerational trauma transmission [13-15] which based on this study would likely lead to elevated risk factors before a child is even born.

References

[6] David Skiena. The Data Science Design Manual

[7] George Engel. The Biopsychosocial Model and the Education of Health Professionals. *University of Rochester, department of Psychiatry and Medicine.*
Link: https://nyaspubs.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1749-6632.1978.tb22070.x

[8] Papadimitriou G. The "Biopsychosocial Model": 40 years of application in Psychiatry. *Psychiatriki.* 2017 Apr-Jun; 28(2):107-110. Greek, Modern, English. PMID: 28686557.
Links: 10.22365/jpsych.2017.282.107 https://pubmed.ncbi.nlm.nih.gov/28686557/

[9] Physiopedia. Biopsychosocial Model. *Online Article*
Link: https://www.physio-pedia.com/Biopsychosocial_Model

[10] National Health Service. Conditions: Fibromyalgia. *NHS online resource*
Link: https://www.nhs.uk/conditions/Fibromyalgia/

[11] Peng & Matsui. The Art of Data Science, a guide for anyone who works with data. *Course reading material.* Book.

[12] Neuroscientists Have Followed a Thought as It Moves Through the Brain. *2018 Online Article*
Link: https://www.sciencealert.com/neuroscience-tracking-thoughts-through-brain-prefrontal-cortex-role

[13] Tory DeAngelis. The legacy of trauma. Feature. *American Psychological Association.*
Link: https://www.apa.org/monitor/2019/02/legacy-trauma

[14] Rachel Yehuda. Intergenerational transmission of trauma effects: Putative role of epigenetic mechanisms. *World Psychiatry Journal Special Article.*
Link: https://onlinelibrary.wiley.com/doi/full/10.1002/wps.20568

[15] Fitzgerald et al. Intergenerational transmission of trauma and family systems theory: An empirical investigation. *Wiley Online Library*
Link: https://www.onlinelibrary.wiley.com/doi/full/10.1111/1467-6427.12303