Computational inference and prediction in public health

Computational inference and prediction in public health

By Steve Bicko Cygu, B.Sc, M.Sc, M.Sc

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

September 12, 2022

McMaster University
Doctor of Philosophy (2022)
Hamilton, Ontario (Computational Science & Engineering)

TITLE: Computational inference and prediction in public health
AUTHOR: Steve Bicko Cygu (McMaster University)
SUPERVISOR: Professor Jonathan Dushoff
NUMBER OF PAGES: xxii, 127

Abstract

Using computational approaches utilizing large datasets to investigate public health information is an important mechanism for institutions seeking to identify strategies for improving public health. The art in computational approaches, for example in health research, is managing the trade-offs between the two perspectives: first, inference and second, prediction. Many techniques from statistical methods (SM) and machine learning (ML) may, in principle, be used for both perspectives. However, SM has a well established focus on inference by building probabilistic models which allows us to determine a quantitative measure of confidence about the magnitude of the effect. Simulation-based validation approaches can be used in conjunction with SM to explicitly verify assumptions and redefine the specified model, if necessary. On the other hand, ML uses general-purpose algorithms to find patterns that best predict the outcome and makes minimal assumptions about the data-generating process; and may be more effective in a number of situations. My work employs both SM- and ML- based computational approaches to investigate particular public health problems. Chapter One provides philosophical background and compares the application of the two approaches in public health. Chapter Two describes and implements penalized Cox proportional hazard models for time-varying covariates time-to-event data. Chapter Three applies traditional survival models and machine learning algorithms to predict survival times of cancer patients, while incorporating the information about the time-varying covariates. Chapter Four discusses and implements various approaches for computing predictions and effects for generalized linear (mixed) models. Finally, Chapter Five implements and compares various statistical models for handling univariate and multivariate binary outcomes for water, sanitation and hygiene (WaSH) data.

Acknowledgements

First and foremost, I would like to thank the Almighty God for His guidance and infinitely abundant grace upon me throughout this research period; He has been a great source of strength.

Next, I would like to express my deepest and profound gratitude to my supervisors, Jonathan Dushoff and Ben M. Bolker, for their guidance and enthusiastic encouragement throughout my graduate studies. Thank you for the guidance beyond my studies. Most importantly, thank you for guiding me to become an independent researcher and helping me realize science's intricate art of communication. Thank you also for showing me how to selflessly and passionately help and support others to advance science. I thank Jonathan for taking a huge gamble to take me as a PhD student and, since then, supported me throughout my PhD journey. I sincerely thank you for offering me this opportunity, and I cannot thank you enough for caring so deeply about my growth as an advisor and a mentor. Thank you for introducing me to the art of reproducible research through "make" and "makestuff" and for your unique way of showing me the power of "make". Witnessing your own dedication, love for science, and desire to make the world a better place was very instrumental in my journey. I thank Ben for his tirelessness and zeal in helping me through this journey, especially when I had trouble with my codes. Thank you for teaching and showing me how to translate my research problems into codes efficiently, and your passion and ability to help others in their research work is something I greatly admire.

I was fortunate to have James Reilly and Hsien Seow on my supervisory committee. Their insightful questions, comments, feedback, suggestions, and recommendations significantly help improve the quality of this thesis. I am very grateful to Hsien for the opportunity to work on PROVIEW project and connecting me with analysts at ICES (formerly the Institute for Clinical Evaluative Sciences) who helped me get the right data and provided useful resources and feedback.

I would also like to thank my collaborators at African Population and Health Research Centre (APHRC) for inviting me to collaborate on Water Sanitation and Hygiene project. I am grateful to Damazo for ensuring I always got access

to data and for the endless support, feedback, and opportunity to network with other researchers in public health.

Sincere thanks to all wonderful past and present members of the MacTheo-Bio Lab: David Champredon, Morgan Kain, Michael Agronah, Greg Forkutza, Jennifer Freeman, Martin Stelmach, Kala Studens, Daniel Park, Steve Walker, Michael Li, and Sana Jahedi. Special thanks to Morgan and Jo Werba for helping me settle in Hamilton; and Michael Li for always being there to listen to me, giving feedback, reminding me about important milestones, and supporting me throughout the PhD journey. Also, a special thanks to Chyun Shi for ensuring a homely and friendly environment in the lab, and of course, not forgetting lab events, especially the pre-COVID-19 delightful dinners.

Lastly, I would like to thank my friends and family members who have not only borne the distance but have been emotionally and spiritually with me. During this journey, people come but mostly give up on me and go. For that, I give special thanks to these individuals: my granfather Owiti Yewa, Dickson & Rhoda Nyambori, Philip Odhiambo, Hezron Osodo, Risper Sigu, Agnes & Brian Kiprop, Samuel Irungu, Fanuel Otieno, Sheila Okwiri, Kogot Peter, Malachi Owino, Catherine Naserian, Collins Onyango, Jackline Kagwiria, Hamidu Mbonde, Mark Kamukama and Brenda Robi, who always believe in me. Special thanks to Brenda Robi for the support and encouragement whenever I felt like giving up.

Dedicated to my late grandmother, Peres Atieno (Nya-Kagwa), who believed in me and without whose unconditional sacrifices and love I would not be where I am today.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Steve Bicko Cygu, declare that this sandwich thesis titled, "Computational inference and prediction in public health" contains:

- An introduction (Chapter 1).
- One submitted manuscript in revision (Chapter 2).
- One submitted manuscript (Chapter 3).
- One manuscript ready for submission (Chapter 4).
- One manuscript in preparation for submission (Chapter 5).
- A conclusion (Chapter 6).

Preambles to Chapters 2-5 describe the authors' contributions. Chapters 1 and 6 were writen by me.

Chapter 1

Introduction

1.1 Background

Public health aims at promoting and improving population health at the individual, community, national, or regional level (Marathe and Ramakrishnan 2013; Santos et al. 2019; Wiemken and Kelley 2020) through various interventions such as prevention of existing and emerging diseases, promotion of healthy lifestyles, timely detection and response to emerging infectious diseases. The goal is to keep the population healthy and ensure prolonged life (Wiemken and Kelley 2020; Bzdok et al. 2020). The need for observing public health has been highlighted in academic research and mainstream media. For instance, the recent outbreak of COVID-19 has been widely publicized to alert the population and advise on appropriate measures to be taken to keep the public healthy and safe (Ahmad et al. 2021).

Emergence and re-emergence of diseases continue to pose challenges to public health researchers in preparing for potential disasters and building a strong response framework in awake of such occurrences. Developing and providing computational techniques to study the dynamics of public health problems can go a long way in helping public health researchers in preparing for the outbreaks and facilitating decision making and policy formulation (Santos et al. 2019; Wiemken and Kelley 2020; Ahmad et al. 2021). In public health, however, deciding on appropriate computational tools and approaches can be challenging for several reasons. For instance, time sensitivity, variability, uncertainty, and a large amount of information associated with public health data (Mhasawade et al. 2020). Nevertheless, computational approaches are increasingly becoming crucial in solving public health-related problems.

Technological advancements and innovations have made it easy for organizations to collect electronic data in public and private institutions. Consequently, there has been incredible growth in the size of data in the public health sector and other domains, such as social media, healthcare, bioinformatics, image processing, search engines, and so on. For example, rapid advancement in genomics technologies combined with the low cost of sequencing has led to the generation of extensive amounts of cancer genomic data (Cagan and Meyer 2017b). This growth presents challenges for effective and efficient data handling and analysis. Such challenges may include data sets with more predictors than number of observations ($p > n$) or several redundant or irrelevant predictors. The term *big data* has been widely used in the literature to describe this kind of data (Marathe and Ramakrishnan 2013; Santos et al. 2019; Dash et al. 2019; Bzdok et al. 2020). Using computational approaches utilizing big datasets to investigate public health problems is an important mechanism for institutions seeking to identify strategies for improving public health – it is a step towards advocating for strategies for improving the population health (Schaik et al. 2019).

The art in computational approaches, for example, in health research, is managing the trade-offs between two perspectives: first, *inference* and second, *prediction*. Prediction aims at forecasting unobserved outcomes from a set of predictors. It makes it possible to establish the best courses of action (e.g., cancer therapy choice) without necessarily requiring an understanding of the underlying process. On the other hand, inference generates statistical, mathematical (or both) models of the data-generation process to formalize understanding or test hypotheses about how the system behaves. In public health research, the two perspectives are not mutually exclusive (Bzdok et al. 2018; Schaik et al. 2019). For example, one may: 1) use a single statistical tool for the two perspectives; or 2) want to infer which biological processes are associated with cancer mortality and predict a cancer diagnosis or prognosis from measured biomarkers.

Take, for example, a simple linear regression, extensively used by public health investigators – the same model can be used for two very different goals: 1) inference, to investigate the impact of a particular input variable on the outcome variable beyond chance, and 2) prediction, to enable us to generalize how well the outcome variable can be guessed from the "unseen" data. The first scenario aims to mechanistically describe the aspects of the inner workings of a phenomenon under investigation, while the second scenario aims to make accurate predictions of future observations.

The two scenarios in the linear model example above differ in terms of motivation, but the mathematical formulation underlying the parameter estimation is, in most cases, equivalent. We would argue that the underlying scientific investigation's desired goal should advise how analysis tools are chosen, and how they are used. Inference-based questions, as mentioned above, are more focused on statistical association and significance testing by examining the effect of the individual input variables on the outcome variable. This is most common in controlled experimental design studies and may involve formulating a null hypothesis (Wasserstein and Lazar 2016; Amrhein et al. 2017; Szucs and Ioannidis 2017).

Much published public health research that uses inference-based approaches relies heavily on null-hypothesis testing. This involves rejecting or failing to reject a null hypothesis of no difference, association or effect depending on the relationship or effect being investigated, based on a p-value cutoff. There are some limitations to this approach: first, it does not give a clear indication of the magnitude of the statistical differences, associations, or effect sizes; and second, the statistical inference becomes more of a binary ("yes" or "no") decision making process (Amrhein et al. 2017; Ioannidis 2018; Dushoff et al. 2019). The use of hypothesis testing may be preferable in some cases. For example, if the scientific goal is to determine which predictors have the most impact on the outcome, hypothesis testing is ultimately a good choice.

As an alternative way to overcome some limitations involving hypothesis testing, there has been growing attention focusing on using confidence intervals (CIs) in public health research (D. Redelings et al. 2012). CIs provide a way to assess and report the precision of a point estimate, such as the proportion of households with access to clean water, average household income, or parameter estimates from a linear model. They provide a mechanism to quantify the degree of uncertainty around a point estimate due to sampling variation at a particular confidence level (usually 95%). In other words, CIs describe how the point estimate could differ in the presence of a "different dataset" if any other underlying condition remained the same.

Traditionally, most public health empirical research has focused on using traditional inferential approaches to, for example, assess the evidence of interventions. These approaches are important for answering questions such as "which prognosis factors are strongly associated with high mortality in cancer patients?" or "what social, economic, and demographic factors contribute most to improved water services among slum dwellers?".

In public health studies, traditional statistical techniques based on generalized linear models are commonly used for drawing conclusions underlying the scientific inquiry (Bzdok et al. 2018; Bzdok and Ioannidis 2019; Bzdok et al. 2020). These techniques' performance and suitability are data-focused and make various assumptions about the data and the model used. Further, as problems become more challenging with advanced research questions, it is becoming more difficult to meet traditional approaches' assumptions, partly due to complex and non-linear relationships between variables, and hence, the need for advanced methods – *prediction* approaches. Flexible prediction approaches are particularly well suited for summarizing these potentially rich datasets to discern patterns (Efron and Hastie 2016; Bzdok et al. 2020).

Prediction approaches provide a way to verify whether the model-derived relationships *predict* "unseen" data points (Bzdok et al. 2020). For example, predictive machine learning algorithms can derive the survival time of new cancer patients (whose survival outcome is not yet observed) based on clinical and demographic information. Prediction-based approaches may be less transparent but have the potential to generate patient-specific predictions in a very fast and effective way. Patient-specific predictions can be used to develop individualized treatment options (Katzman et al. 2018a; Paulus and Thompson 2021).

Moreover, prior studies have shown the potential of prediction approaches, especially machine learning algorithms in public health and biomedical data (Hinton and Salakhutdinov 2006; Esteva et al. 2017; Poplin et al. 2018; Hannun et al. 2019). Despite this huge potential, the empirical success of predictive models may not fully satisfy the scientific curiosity underlying the investigation to provide an understanding of the research problem.

In prediction, the purpose is to search through all the possible meaningful patterns in the data to extract knowledge about regularities (Bzdok et al. 2018; Bzdok and Ioannidis 2019; Bzdok et al. 2020). This approach is, for example, suited for questions like "can we use available predictors to accurately predict cancer reccurrence?". The success of predictive models is usually quantified using a prediction accuracy metric. In some cases, the discovered relationships may be opaque and not accessible to the investigator – "black box". Prediction approaches use external validation to improve the prediction accuracy of the trained model in identifying the outcomes of new observations in the "unseen" data.

In public health research, like any other field, investigators have had different perspectives concerning the best ways to analyze and handle new kinds of data sources, especially for big data. Traditional inference-based approaches mentioned above were introduced and predominantly used when the data was scarce, or there was limited access to the data, and have been revised, or even in some cases, researchers advised to use them with a lot of caution (Amrhein et al. 2017; Ioannidis 2018). There has been a growing literature promoting the use of prediction-based algorithms capable of providing quick insights into the big and complex datasets (Bzdok and Ioannidis 2019; Santos et al. 2019; Wiemken and Kelley 2020; Ahmad et al. 2021). Such predictive models are gaining momentum in public health and other biomedical fields (Jordan and Mitchell 2015; LeCun et al. 2015; Bzdok and Ioannidis 2019). However, it might not be easy to distinctly categorize a particular analysis tool into a particular category like "statistics" or "machine-learning" (Bzdok and Ioannidis 2019).

The need to improve predictive performance may necessitate the use of more complicated prediction-based models as opposed to the widely used inference-based, arguably more transparent, approaches such as linear models, the test of association, and hypothesis testing, which have been widely used in public health studies (Visscher et al. 2017; Bzdok and Ioannidis 2019). The data-driven predictive models aimed at identifying non-linear relationships between variables have a strong legacy in machine learning methods and have continued to gain recognition in public health over recent years (Efron and Hastie 2016; LeCun et al. 2015; Bzdok and Ioannidis 2019).

In principle, many statistical methods (SM) and machine learning (ML) techniques may be used for both perspectives. However, SM has a well-established focus on inference by building probabilistic models, which allow us to determine a quantitative measure of confidence about the magnitude of the effect. Simulation-based validation approaches can be used with SM to verify assumptions and redefine the specified model, if necessary. On the other hand, ML uses general-purpose algorithms to find patterns that best predict the outcome and makes minimal assumptions about the data-generating process; and may be more effective in several situations, for example; 1) where the number of predictors exceeds the number of observations (wide data), 2) high-dimensional data with high storage and computational requirements, and 3) in the presence of complicated non-linear interactions (Bzdok et al. 2018). However, despite convincing predictive performance and flexibility, the lack of explicit models in most ML methods can make ML results difficult to link to prior public health knowledge directly.

1.2 Research Aim

This research employs both SM- and ML- based computational approaches to investigate particular public health problems. The former primarily uses simulation to validate and refine our assumptions and make causal inferences about the estimated model parameters. The latter uses cross-validation to evaluate models for maximized predictive performance on unobserved outcomes. I focus on two specific objectives:

- To build and compare traditional and machine learning methods to predict survival times for cancer patients.
- To investigate the contribution of demographic and economic factors to improved water, toilet facilities, and garbage collection among the Nairobi urban poor using multivariate multilevel models for binary outcomes.

1.3 Goals

The following are some of the intended contributions of this research to the existing literature:

1. To discuss, extend and implement penalized Cox proportional hazard models to handle time-varying covariate time-to-event data.
2. To build and compare traditional and hazard-based machine learning models that can be used to predict survival times of cancer patients while incorporating the information about time-varying covariates.
3. To extend and implement frameworks for summarizing and visualizing predictions and effects for generalized linear (mixed) models.
4. To provide an alternative method for bias correction for predictions and effects for generalized linear (mixed) models.
5. To provide a more in-depth understanding of the public health data in the context of water, sanitation, and hygiene (WaSH) and extend the coverage of computational approaches to these different types of data.
6. To compare various statistical models for handling univariate and multivariate binary outcomes for WaSH data.

1.4 Thesis Outline

The remainder of the thesis is organized as follows:

In **Chapter 2**, I describe and implement an algorithm and R package (**pcoxtime**) for penalized Cox proportional hazard (CPH) models with time-dependent covariates. Until recently, the standard R packages for fully penalized Cox models could not incorporate time-dependent covariates. To address this gap, I implemented a *proximal gradient descent* algorithm for fitting penalized Cox models and applied it to real and simulated data sets.

In **Chapter 3**, I focus on the application of machine learning algorithms to cancer data. Using data for adults diagnosed with cancer from a population-based, retrospective study collected from January 2008 to December 2015 from Ontario, Canada, I build, validate, and compare traditional CPH models and CPH-based machine learning models for both time-invariant and time-varying covariates and further compare the performance of these models to a prior analysis implemented by Seow et al. (2020).

In **Chapter 4**, I discuss and implement various approaches for computing predictions and effects. I further explore and demonstrate approaches for correcting bias in the central estimates for generalized linear (mixed) models involving non-linear link functions. I use simulation to illustrate two (mean-based and observed-value-based) approaches for generating predictions and effects, and show that they can produce substantially different results.

In **Chapter 5**, I describe, develop, perform simulation-based validation and apply a joint modeling approach to analyze binary outcomes. The approach takes into account the longitudinal nature of the data to model all the three water, sanitation, and hygiene (WaSH) outcome variables (improved water, toilet facilities, and garbage disposal). The analysis is based on a generalized linear mixed model approach; and compares separate (univariate) and joint (multivariate) models for binary outcomes to investigate the contribution of demographic and socio-economic factors to WaSH access in two informal urban settlements in Nairobi, Kenya, namely Korogocho and Viwandani.

Chapter 2

Penalized Cox Proportional Hazard Model for Time-dependent Covariates

The text I present here is a manuscript already submitted for publication and is in revision

Author Contributions

SC performed statistical analyses with helpful feedback from JD and BMB; SC wrote the first draft of the manuscript, and all authors revised the manuscript.

pcoxtime: Penalized Cox Proportional Hazard Model for Time-dependent Covariates

Steve Cygu[1,⋆], Jonathan Dushoff[1,2], Benjamin M. Bolker[1,2]

[1]McMaster University, School of Computational Science and Engineering, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[2]McMaster University, Department of Biology, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

⋆cygubicko@gmail.com

Abstract

The penalized Cox proportional hazard model is a popular analytical approach for survival data with a large number of covariates. Such problems are especially challenging when covariates vary over follow-up time (i.e., the covariates are time-dependent). The standard R packages for fully penalized Cox models cannot currently incorporate time-dependent covariates. To address this gap, we implement a variant of gradient descent algorithm (*proximal gradient descent*) for fitting penalized Cox models. We apply our implementation to real and simulated data sets.

Keywords: survival, time-dependent, Cox proportional hazard, elastic net, penalized, proximal

2.1 Introduction

Survival analysis studies event times, such as time to cancer recurrence or time to death. Its goal is to predict the time-to-event (survival time) using a set of covariates, and to estimate the effect of different covariates on survival. Survival models typically attempt to estimate the *hazard*, the probability (density) of the occurrence of the event of interest within a specific small time interval. Binary classification methods from machine learning can be used in problems that focus on predicting whether an event occurs within a specified time window. However, while binary classifiers can predict outcomes for a specified time window, they fail to account for one of the unique characteristics of survival data — *censoring*. In survival data, some of the subjects may be lost to follow-up, or may be event-free by the end of the follow-up time; hence the event times represent censoring times rather than failure (death, recurrence, etc.) times. Since binary classifiers consider only whether or not the event occurred in the last observation window, they lack the interpretability and flexibility of models that consider hazards as a function of time (Kvamme et al. 2019).

Cox proportional hazard (CPH) models are the most common approach in survival analysis. Traditionally, the CPH model has been applied in problems where the number of observations, n, is much larger than the number of covariates, p. In the modern era of big data, however, researchers often encounter cases where $p \approx n$ (or $p \gg n$). In cancer research, for example, rapid advances in genomic technologies have led to the generation of vast amounts of cancer data (Cagan and Meyer 2017a) — presenting inherent challenges for effective and efficient data analysis. Penalized regression methods such as *lasso*, *ridge* or *elastic net* offer a statistically convenient way of handling high-dimensional data, especially when building predictive models. The subclass of penalized methods which are sparsity-inducing (e.g., lasso and elastic net) can also be used to select useful predictive features from a large set.

The standard CPH model (i.e., with no time-dependent covariates) assumes that the hazard ratio is constant over the entire follow-up period, or equivalently that each covariate is fixed over time and has a constant multiplicative effect on the hazard function. This assumption is problematic when covariates of interest themselves change over time. For example, cancer patients' healthcare access may change over the course of a study. Some implementations of CPH models allow such *time-dependent covariates*. However, their use requires more attention than the fixed (time-independent) covariates (Hochstein et al. 2013; Therneau et al. 2017; Austin et al. 2020).

Many authors have implemented CPH models with penalization, but many implementations (Gui and Li 2005; Park and Hastie 2007; Sohn et al. 2009; Goeman 2010) are computationally inefficient, due to their use of the Newton-Raphson algorithm (Gorst-Rasmussen and Scheike 2012). Some newer implementations are more efficient: Simon et al. (2011) describe and implement an impressively fast algorithm **coxnet**, implemented in the **glmnet** package, for fitting regularized CPH models via weighted cyclic coordinate descent. This method is computationally efficient in handling high-dimensional problems. Yang and Zou (2013) proposed and implemented the **cocktail** algorithm, which is a mixture of coordinate descent, the majorization-minimization principle, and the strong rule for solving penalized CPH models in high dimensional data. The **cocktail** algorithm (implemented in the **fastcox** package) always converges to the correct solution and is slightly faster than the **coxnet** algorithm. However, these implementations, the benchmark R packages for penalized Cox models, have some limitations. The implementations by Simon et al. 2011 and Yang and Zou 2013 do not support time-dependent covariates; the implementation by Goeman 2010 does incorporate time-dependent covariates, but only implements *naive* elastic net, neglecting subsequent improvements in the algorithm (Simon et al. 2011).

Other, non-CPH-based, approaches have also incorporated time-dependent covariates in penalized models for time-to-event-data. Most such approaches have used generalized additive models to implement semiparametric regression methods in the context of survival models (Gorst-Rasmussen and Scheike 2012; Bender et al. 2018). Gorst-Rasmussen and Scheike 2012 used a cyclic coordinate descent algorithm to develop a penalized semiparametric additive hazard model (in the **ahaz** package). The model defines a hazard function as the sum of the baseline hazard and the regression function of the covariates — it is intrinsically linear thus theoretically guarantees convergence, and can handle time-dependent covariates. However, currently, it only implements lasso penalization.

In this paper, we describe and implement an algorithm and R package (**pcoxtime**) for penalized CPH models with time-dependent covariates. The general properties of penalized methods make this algorithm a useful tool for handling high-dimensional problems. We describe how existing computational approaches for CPH modeling can be adapted to obtain penalized methods for time-dependent covariates in time-to-event data. To solve the optimization problem, we exploit a variant of the gradient descent algorithm known as proximal gradient descent (as outlined in Parikh and Boyd 2014) with Barzilai-Borwein step-size adjustment (Barzilai and Borwein 1988). Unfortunately, the gradient-descent approach here is intrinsically slower than methods based on coordinate descent (Simon et al. 2011); we are working to implement coordinate descent. In the meantime, the capabilities and convenience of **pcoxtime** will still be useful for moderately large problems.

We test our package on simulated data with time-dependent covariates, and compare its performance with that of the **penalized**. We also provide examples of its usage on real data.

2.2 Methods and algorithms

2.2.1 Cox model with time-independent covariates

Survival data is often presented in the form $\{t_i, \delta_i, x_i\}_{i=1}^n$, where t_i is the observed event time (failure time or censoring time) for individual i, δ_i is an indicator variable for whether the observed endpoint is a failure (rather than censoring), and x_i is a vector of covariates $(x_{i,1}, x_{i,2}, \cdots, x_{i,p})$.

The CPH model (Cox 1972) defines the hazard function at time t as

$$h_i(t) = h_0(t) \exp{(x_i^\top \beta)}, \tag{2.1}$$

where $h_0(t)$ is the non-parametric baseline hazard and β is the coefficient vector of length p.

In a simple case where there are no ties, with $t_1 < t_2 < \cdots < t_k$ representing unique ordered event (or failure) times, we can define the *risk set* R_i, of individuals who are still at risk of failing (not yet censored or failed) at time t_i – individuals with event time $t_j \geq t_i$. The likelihood function corresponding to the order of events (Simon et al. 2011; Yang and Zou 2013) is given by

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\exp\left(x_i^\top \beta\right)}{\sum_{j\in R_i} \exp\left(x_j^\top \beta\right)}, \tag{2.2}$$

and we can thus optimize the parameters β by maximizing the partial log-likelihood:

$$\ell(\beta) = \sum_{i:\delta_i=1} \left(x_i^\top \beta - \log \left[\sum_{j\in R_i} \exp\left(x_j^\top \beta\right) \right] \right). \tag{2.3}$$

The Cox model in Equation 3.1 is fitted in two steps — first, the parametric part is fitted by maximizing the partial log-likelihood in Equation 2.3, and then the non-parametric baseline hazard is estimated.

Following the **survival** (Therneau 2020) package, here we use a slightly more general formulation, where the observed survival data is of the form $\{t_i^{\text{start}}, t_i^{\text{stop}}, \delta_i, x_i\}_{i=1}^n$, where t_i^{start} and t_i^{stop} bracket the period in which the event time for the i^{th} individual occurred. This formulation allows for greater flexibility in defining the time scale on which the analysis is based (e.g., time since diagnosis vs. time of followup); it will also allow us to address left censorship in the observation of outcomes and (later) in the observation of covariates. The risk set at time t_i is now defined as $R_i(t) = \{j : (t_j^{\text{start}} < t_i) \cap (t_i \leq t_j^{\text{stop}})\}$. The first condition, $(t_j^{\text{start}} < t_i)$, ensures the start time was observed before the event, while the second condition, $(t_i \leq t_j^{\text{stop}})$, ensures that individual j either experienced the event or was censored at a later time point than t_i.

2.2.2 Cox model with time-dependent covariates

When a covariate changes over time during the follow-up period, the observed survival data is of the form $\{t_i^{\text{start}}, t_i^{\text{stop}}, \delta_i, x_i(t)\}_{i=1}^n$. The only difference is that x_i is now a (piecewise constant) function of time. Using Breslow's approximation (Breslow 1972) for tied events, the partial log-likelihood is defined as

$$\ell(\beta) = \sum_{i=1}^{k} \left(\left[\sum_{s\in D_i} x_s^\top(t)\beta \right] - d_i \log \left[\sum_{j\in R_i(t)} \exp(x_j^\top(t)\beta) \right] \right), \tag{2.4}$$

where d_i is the number of failures at time t_i, $k < n$ (if there are ties), D_i are the set of indexes j for subjects failing at time t_i and the description of $t_1, t_2, \cdots, t_k$ remains the same as in the previous case (Harrell Jr 2015). The parameter estimates $\hat{\beta}$ are obtained by minimizing $-\ell(\beta)$.

2.2.3 Algorithm

Our algorithm extends the partial log-likelihood in Equation 2.4 by adding the penalty term. We let $P_{\alpha,\lambda}(\beta)$ be a mixture of ℓ_1 (lasso) and ℓ_2 (ridge) penalties. The penalized Cox partial log-likelihood (objective function) is defined as

$$\Omega(\beta)_{\alpha,\lambda} = -\ell(\beta) + P_{\alpha,\lambda}(\beta),$$

where

$$P_{\alpha,\lambda}(\beta) = \lambda \left(\alpha \sum_{i=1}^{p} |\beta_i| + 0.5(1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right) \tag{2.5}$$

as proposed by Zou and Hastie 2005, with $\lambda > 0$ and $0 \leq \alpha \leq 1$. The lasso penalty $(\sum_{i=1}^{p} |\beta_i|)$ induces sparsity by selecting a subset of nonzero coefficients. It works well in high-dimensional applications, but will eliminate all but one of any set of strongly multicollinear terms. On the other hand, the ridge penalty $(\sum_{i=1}^{p} \beta_i^2)$ shrinks coefficients towards but never all the way to zero; hence it gives non-sparse estimates and can give correlated predictors approximately equal weights. The elastic net penalty combines the strength of lasso and ridge penalties for improved predictive performance (Simon et al. 2011). As α increases, the sparsity and the magnitude of non-zero coefficients decreases, i.e., the solution becomes less ridge-like and more lasso-like.

Using Equation 2.5, our minimization problem becomes

$$\hat{\beta} = \arg\min_{\beta} \;\; \Omega(\beta)_{\alpha.\lambda}. \tag{2.6}$$

Parameter estimation

The lasso penalty is not differentiable at $\beta = 0$. We thus solve the minimization problem above using *proximal gradient descent* by decomposing Equation 2.6, the objective function, as $f(\beta) = g(\beta) + h(\beta)$, with

$$g(\beta) = -\ell(\beta) + 0.5\lambda(1-\alpha)\sum_{i=1}^{p}\beta_i^2$$

and

$$h(\beta) = \lambda\alpha\sum_{i=1}^{p}|\beta_i|.$$

In this form, we split the objective function, $\arg\min_{\beta}\ \Omega(\beta)_{\alpha,\lambda}$, into two parts, one of which is differentiable. Specifically, $g(\beta)$ is differentiable and convex and $h(\beta)$ is convex but not necessarily differentiable. The proximal gradient operator (Parikh and Boyd 2014) to update β is given by

$$\beta^{(k)} = \text{prox}_{\gamma_k h}\left(\beta^{(k-1)} - \gamma_k\nabla g(\beta^{(k-1)})\right),\ k = 1, 2, 3, \cdots \tag{2.7}$$

where γ_k is the step size determined via Barzilai-Borwein step-size adjustment (Barzilai and Borwein 1988). Park and Hastie 2007 shows that $\text{prox}_{\gamma_k h}(.)$ reduces to (elementwise) *soft thresholding*

$$\text{prox}_{\gamma_k\lambda\alpha}(x_i) = \begin{cases} x_i - \gamma_k\lambda\alpha & x_i \geq \gamma_k\lambda\alpha \\ 0 & -\gamma_k\lambda\alpha \leq x_i \leq \gamma_k\lambda\alpha \\ x_i + \gamma_k\lambda\alpha & x_i \leq -\gamma_k\lambda\alpha \end{cases} \tag{2.8}$$

and

$$\begin{aligned}\nabla g(\beta) &= -\nabla\ell(\beta) + \lambda(1-\alpha)\sum_{i=1}^{p}\beta_i \\ &= -\sum_{i=1}^{k}\left(\sum_{s\in D_i} x_s^\top(t) - d_i\frac{\sum_{j\in R_i(t)} x_j^\top(t)\exp(x_j^\top(t)\beta)}{\sum_{j\in R_i(t)}\exp(x_j^\top(t)\beta)}\right) + \lambda(1-\alpha)\sum_{i=1}^{p}\beta_i \\ &= \pi(\beta) + \lambda(1-\alpha)\sum_{i=1}^{p}\beta_i. \end{aligned} \tag{2.9}$$

Our package implements the Karush–Kuhn–Tucker (KKT) conditions check described in Yang and Zou 2013 to test that the β estimates are valid. We did not come across any convergence problems in the examples analyzed here (i.e., the KKT conditions were always satisfied)

To train an optimal model, we need to choose a value of λ. With a large value of λ the penalty terms in Equation 2.6 will dominate, driving coefficients to zero, while a small λ value will lead to overfitting. We can use cross-validation to pick an optimal λ from a set of λ values (known as a regularization path) $\lambda_1 < \lambda_2, \cdots, < \lambda_{\text{max}}$. We want λ_{max} to be large enough that $\beta = \mathbf{0}$, and λ_1 to be small enough to give a result close to the unpenalized solution (this choice enables the *warm-start* approach employed in **glmnet**).

From Equation 2.8 and Equation 2.9 notice that if $\pi(\beta) \leq \alpha\lambda\gamma_k$, then $\beta^k = 0$ minimizes our objective function. Thus we set λ_{max} to be

$$\lambda_{\text{max}} = \frac{1}{N\alpha\gamma_k} \max_{\beta} \left\{ \sum_{i=1}^{k} \left(\left[\sum_{s \in D_i} x_s^\top(t) \right] - \frac{d_i}{|R_i(t)|} \left[\sum_{j \in R_i(t)} x_j^\top(t) \right] \right) \right\}, \tag{2.10}$$

where $|R_i(t)|$ denotes the cardinality of the risk set $R_i(t)$. If $\alpha = 0$, we set $1/N\alpha\gamma_k$ in Equation 2.10 to $1/0.001N\gamma_k$.

In our implementation, we set $\lambda_{\text{min}} = \epsilon\lambda_{\text{max}}$, and compute solutions over a grid of m values of λ decreasing from λ_{max} to λ_{min}, where $\lambda_i = \lambda_{\text{max}}(\lambda_{\text{min}}/\lambda_{\text{max}})^{i/(m-1)}$ for $i = 0, \cdots, m-1$ (Simon et al. 2011). The default value of k is 100 (the number of distinct λ values). If $n \geq p$, the default value of ϵ is set to 0.0001; otherwise (i.e., if $n < p$), $\epsilon = 0.01$ (Yang and Zou 2013; Simon et al. 2011).

Cross-validation

Most implementations cross-validate over a range of λ values for a fixed α. However, our implementation allows the user to choose a range of α, $0 \leq \alpha \leq 1$; in this case the algorithm will pick the α-λ pair that corresponds to the lowest cross-validated partial likelihood deviance (CV-PLD) or highest cross-validated Harrell's concordance index (CV-C-index) (Harrell Jr et al. 1996).

To find the CV-PLD for each λ-α pair, we perform k-fold cross-validation — the training data is split into k folds, and the model is trained on $k-1$ folds and validated on the left-out part via some predictive performance measure k times. Here, we implement two metrics i.e., CV-PLD and CV-C-index (Dai and Breheny 2019). The CV-PLD is:

$$\text{CV-}\widehat{\text{PLD}}(\lambda) = -2\sum_{k=1}^{K} \ell(\hat{\beta}_{-k}(\lambda)) - \ell_{-k}(\hat{\beta}_{-k}(\lambda)) \tag{2.11}$$

where $\ell(\hat{\beta}_{-k}(\lambda))$ is the log partial likelihood evaluated at $\hat{\beta}_{-k}$ using the whole dataset and $\ell_{-k}(\hat{\beta}_{-k}(\lambda))$ is the log partial likelihood evaluated at $\hat{\beta}_{-k}$ on the retained data (everything except the left-out part). The $\hat{\beta}_{-k}$ values denote the penalized estimates using the retained data. We choose the λ which minimizes Equation 2.11. Note that Equation 2.11 gives different (and usually better) results than simply evaluating the partial likelihood on the held-out set (sometimes called the *basic* approach), because the likelihood of any observation depends on other elements in the risk set.

The alternative cross-validation metric, CV-C-index, uses the concordance statistic for Cox models, known as the cross-validated C-index, based on Harrell's concordance index (Harrell Jr et al. 1996). It computes the probability that, for a random pair of individuals, the predicted survival times of the pair have the same ordering as their true survival times. Our implementation is similar to that of the **survival** package (Therneau 2020).

2.2.4 Prediction

Once $\hat{\beta}$ is estimated, we can estimate the baseline hazard function ($\hat{h}_0(t)$), and hence the survival function ($\hat{S}_0(t)$). We first compute the cumulative hazard function

$$\hat{h}_0(t) = \sum_{i \in y_j < t_i} \frac{\sigma_i}{\sum_{j \in R_i(t)} \exp(x(t)_j^\top \hat{\beta})} \tag{2.12}$$

and then, for a given covariate vector, x_i, the estimated hazard, $\hat{h}(t|x_i)_i$, and survival functions are

$$\hat{h}(t|x_i)_i = \hat{h}_0(t)\exp\left(x_i^\top(t)\hat{\beta}\right)$$
$$\hat{S}(t|x_i)_i = \exp\left(-\hat{h}(t|x_i(t))_i \exp\left(x_i^\top(t)\hat{\beta}\right)\right).$$

2.3 Illustrations

In the following sections, we demonstrate the practical use of **pcoxtime** on real and simulated data sets. In the first two examples, we consider real data sets with time-independent covariates and then a time-dependent covariates. The last example compares **pcoxtime** with **penalized** on a simulated data set with time-dependent covariates.

2.3.1 Time-independent covariates

We use the `sorlie` gene expression data set (Sorlie and Tibshirani 2003), which contains 549 gene expression measurements together with the survival times for 115 females diagnosed with cancer. This data set was also used by Gorst-Rasmussen and Scheike 2012 to demonstrate the performance of the **ahaz** package.

We perform a penalized survival regression by varying both α and λ. If a range of α values is desired, we suggest first running the analysis for an intermediate range of α values. If the minimum cross-validation likelihood deviance (Min. CV-PLD) based on k-fold cross-validation (over all λ values considered) is at the lower bound of the range considered, then extend the range of the α vector to lower (positive) values; if it is at the upper bound, extend the range upward. In this example, we cross-validate several α values at the same time by setting $\alpha = \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$. For each α, we analyze a solution path of λ values and use 10-fold cross validation to choose the optimal (Min. CV-PLD) value of α and λ.

We first load the data as follows:

```
R> data("sorlie", package = "ahaz")
```

It is common practice to standardize the predictors before applying penalized methods. In **pcoxtime**, predictors are scaled internally (but the user can choose to output coefficients on the original scale [the default] or to output standardized coefficients). We make the following call to `pcoxtimecv` to perform 10-fold cross-validation to choose the optimal α and λ.

```
R> cv_fit1 <- pcoxtimecv(Surv(time, status) ~., data = sorlie,
+    alphas = c(0.1, 0.2, 0.4, 0.6, 0.8, 1), lambdas = NULL,
+    devtype = "vv", lamfract = 0.8, refit = TRUE, nclusters = 4
+  )
Progress: Refitting with optimal lambdas...
```

In order to reduce the computation time, we use `lamfract` to set the proportion of λ values, starting from λ_{max}, to 80%. Setting `lamfract` in this way specifies that only a subset of the full sequence of λ values is used (Simon et al. 2011).

Once cross-validation is performed, we can report the λ and α for which CVE attains its minimum ($\lambda = 0.796, \alpha = 0.1$) and view the cross-validated error plot (Figure 2.1) and the regularization path (Figure 2.3).

```
R> print(cv_fit1)
Call:
pcoxtimecv(formula = Surv(time, status) ~ ., data = sorlie, alphas = c(0,
    1), lambdas = NULL, lamfract = 0.8, devtype = "vv", refit = TRUE,
    nclusters = 4)

Optimal parameter values
 lambda.min lambda.1se alpha.optimal
  0.7960954   2.215183           0.1
R>
R> cv_error1 <- plot(cv_fit1, g.col = "black", geom = "line",
+     scales = "free")
R> print(cv_error1)
R>
R> solution_path1 <- (plot(cv_fit1, type = "fit") +
```

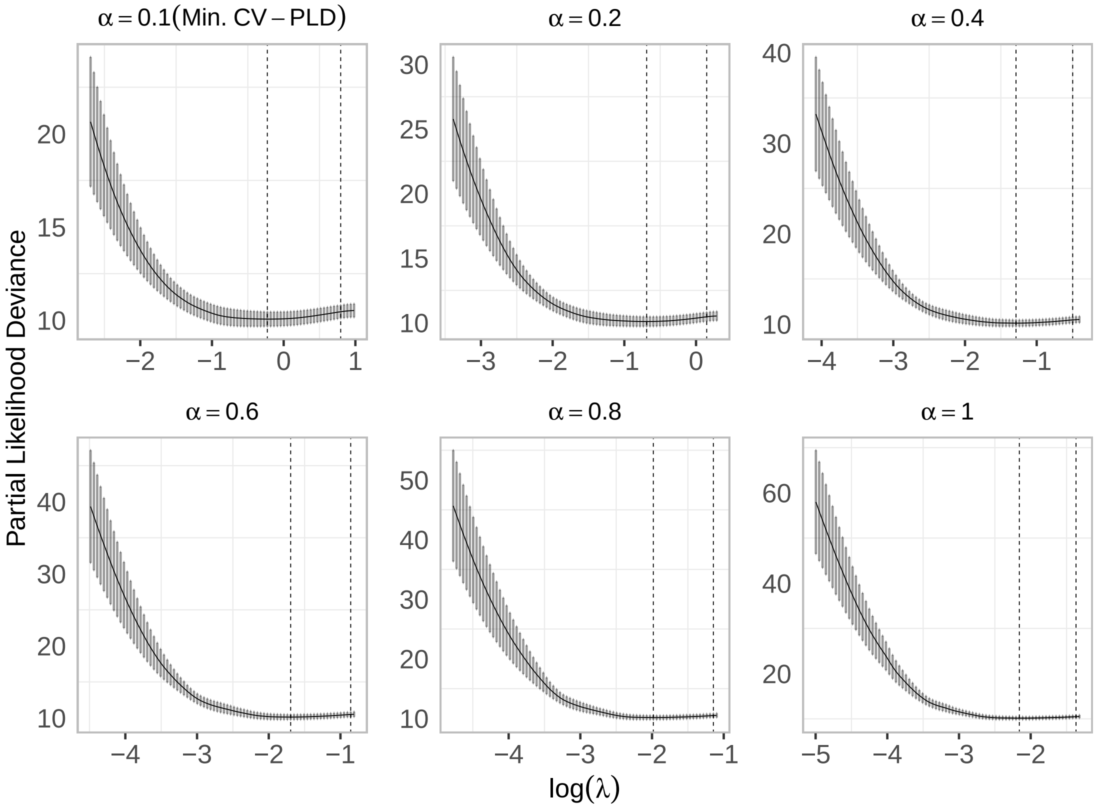


Figure 2.1: Plots of the cross-validated error rates for a sequence of λ values at different α values. If a vector of α values is specified, `pcoxtimecv` automatically chooses the α that minimizes CV-PLD (Min. CV-PLD). In this example, we choose from $\alpha = \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$. The left dotted line indicates the minimum error; the right dotted line indicates the largest value of λ that fits the simplest model whose error is within one standard deviation of the minimum cross-validation error (Hastie et al. 2009). In this case, the best fit occurs when $\alpha = 0.1$ (more ridge-like, top left panel); note that panels differ in their horizontal and vertical scales.

```
+     #ylim(c(-0.03, 0.03)) +
+     labs(caption = "(a) sorlie") +
+     theme(plot.caption = element_text(hjust=0.5, size=rel(1.2)))
+ )
R> print(solution_path1)
```

Next, we fit the penalized model using the optimal α and λ:

```
R> ## Optimal lambda and alpha
```

```
R> alp <- cv_fit1$alpha.optimal
R> lam <- cv_fit1$lambda.min
R>
R> ## Fit penalized cox model
R> fit1 <- pcoxtime(Surv(time, status) ~., data = sorlie,
+    alpha = alp, lambda = lam
+  )
R> print(fit1)
Call:
pcoxtime(formula = Surv(time, status) ~ ., data = sorlie, alpha = alp,
    lambda = lam)

66 out of 549 coefficients are nonzero
n = 115 , number of events = 38
```

We then plot the predicted survival function for each patient and the average survival function (Figure 2.4).

```
R> surv_avg <- pcoxsurvfit(fit1)
R> surv_df <- with(surv_avg, data.frame(time, surv))
R> surv_ind <- pcoxsurvfit(fit1, newdata = sorlie)
R> splot_sorlie <- plot(surv_ind, lsize = 0.05, lcol="grey")
R> splot_sorlie <- (splot_sorlie +
+    geom_line(data = surv_df, aes(x = time, y = surv, group = 1),
+    col = "red") +
+    labs(caption = "(a) sorlie") +
+    theme(plot.caption = element_text(hjust=0.5, size=rel(1.2)))
+ )
R> print(splot_sorlie)
```

2.3.2 Time-dependent covariates

We now repeat the analysis outlined above in the context of survival data with time-dependent covariates. We consider the chronic granulotomous disease (`cgd`) data set from the **survival** package (Therneau 2020), which contains data on time to serious infection for 128 unique patients. Because some patients are observed for more than one time interval, with different covariates in each interval, the data set has 203 total observations.

We load the data and perform cross-validation:

```
R> data("cgd", package = "survival")
R> dat <- cgd
R> cv_fit2 <- pcoxtimecv(Surv(tstart, tstop, status) ~ treat + sex +
+    ns(age,3) + height + weight + inherit + steroids + propylac +
+    hos.cat, data = cgd, alphas = c(0.2, 0.5, 0.8), lambdas = NULL,
+    devtype = "vv", lamfract = 0.6,  refit = TRUE, nclusters = 4
+  )
Progress: Refitting with optimal lambdas...
```

Here, we choose $\alpha = \{0.2, 0.5, 0.8\}$ for 10-fold cross-validation.

```
R> print(cv_fit2)
Call:
pcoxtimecv(formula = Surv(tstart, tstop, status) ~ treat + sex +
    ns(age, 3) + height + weight + inherit + steroids + propylac +
    hos.cat, data = cgd, alphas = c(0.2, 0.5, 0.8), lambdas = NULL,
    lamfract = 0.6, devtype = "vv", refit = TRUE, nclusters = 4)


Optimal parameter values
 lambda.min lambda.1se alpha.optimal
 0.01477729  0.3834742           0.5
```

The Min. CV-PLD (optimal) hyperparameter values are $\lambda = 0.015$ and $\alpha = 0.5$ (Figure 2.2).

```
R> cv_error2 <- plot(cv_fit2, g.col = "black", geom = "line",
+     g.size = 1)
R> print(cv_error2)
```

We plot the solution path (Figure 2.3) and fit the penalized model based on the optimal λ and α:

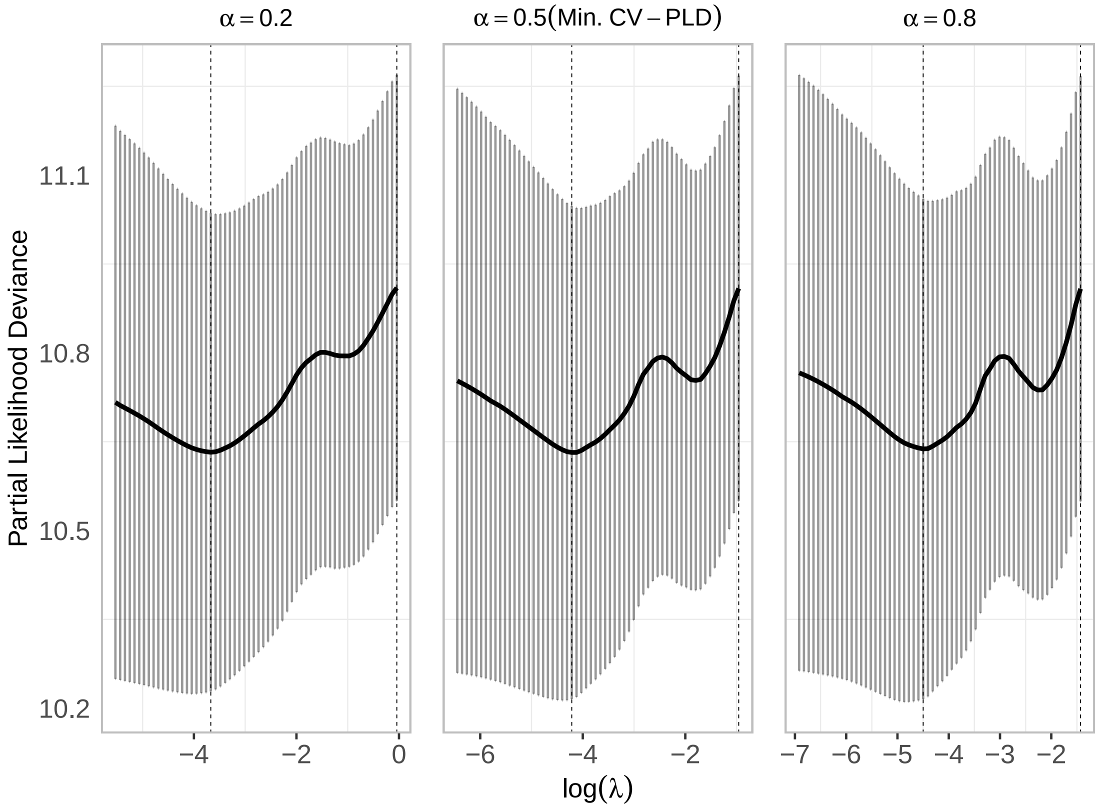
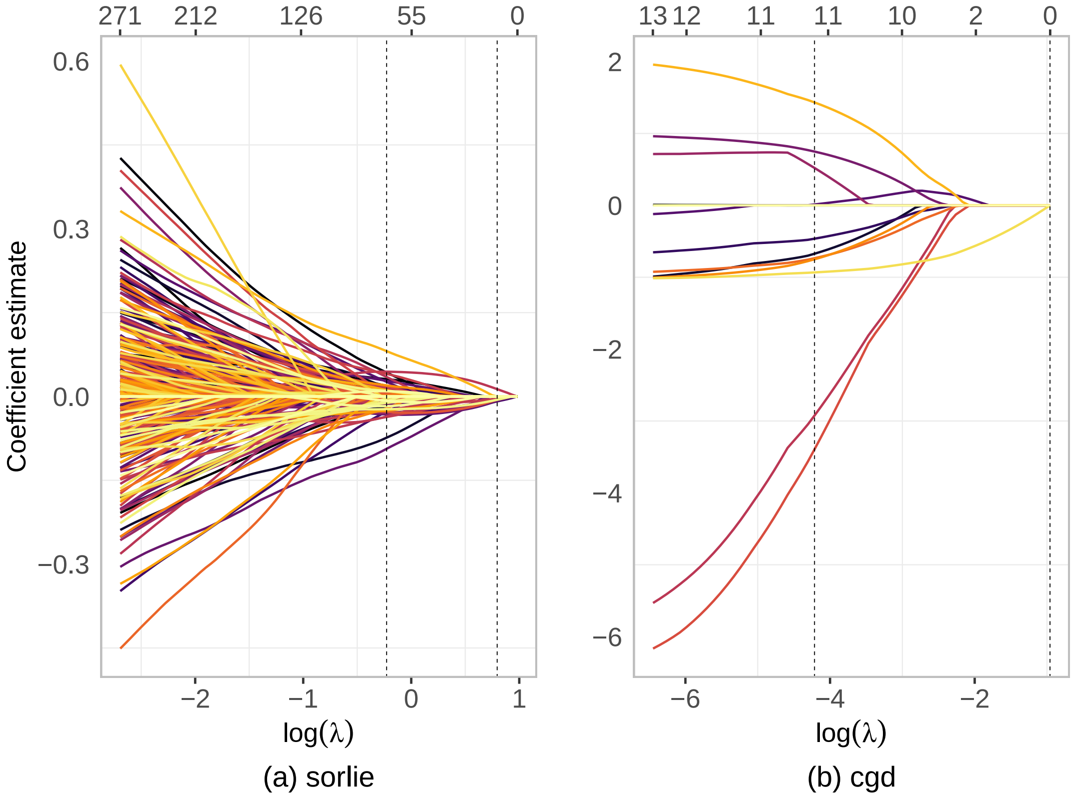


FIGURE 2.2: Cross-validated error rates for the `cgd` data.

```
R> solution_path2 <- (plot(cv_fit2, type = "fit") +
+    labs(caption = "(b) cgd") +
+    theme(plot.caption = element_text(hjust=0.5, size=rel(1.2)))
+ )
R> print(solution_path2)
R>
R> alp <- cv_fit2$alpha.optimal
R> lam <- cv_fit2$lambda.min
R>
R> ## Fit penalized cox model
R> fit2 <- pcoxtime(Surv(tstart, tstop, status) ~ treat + sex +
+    ns(age,3) + height + weight + inherit + steroids + propylac +
+    hos.cat, data = cgd, alpha = alp, lambda = lam
```



(a) sorlie

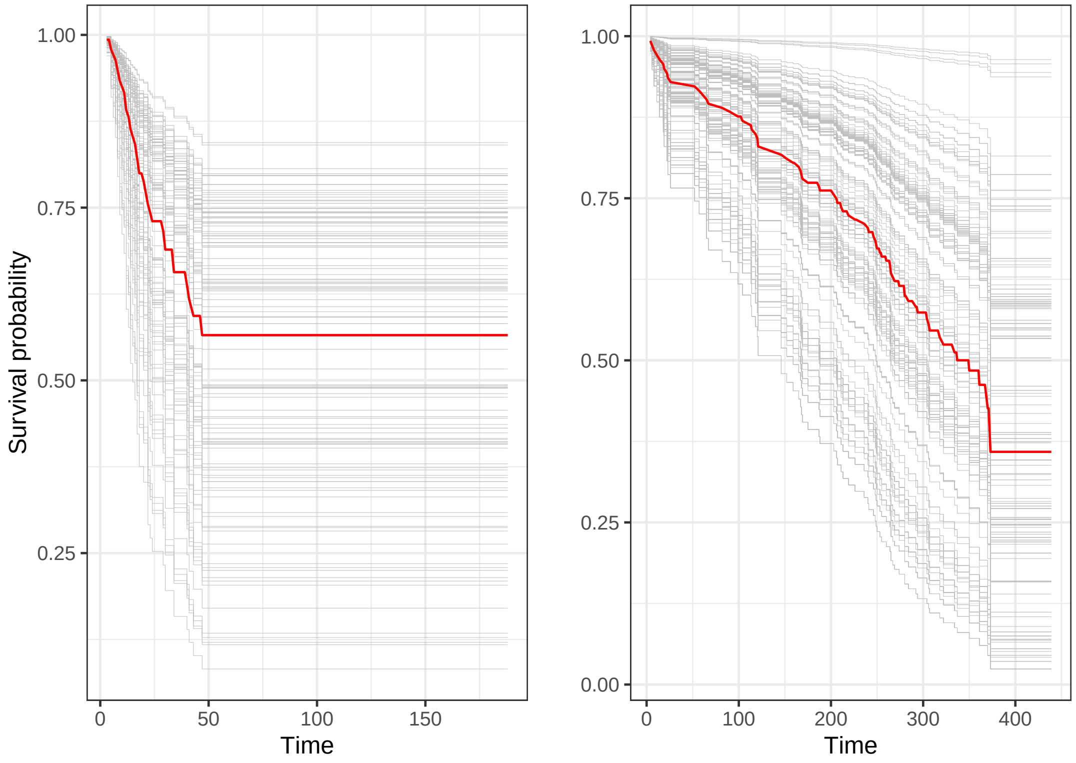
(b) cgd

Figure 2.3: Regularization paths for the `sorlie` and `cgd` models. The values at the top of the plot give the number of nonzero coefficients (size of the model) at various λ values.

```
+  )
R> print(fit2)
Call:
pcoxtime(formula = Surv(tstart, tstop, status) ~ treat + sex +
    ns(age, 3) + height + weight + inherit + steroids + propylac +
    hos.cat, data = cgd, alpha = alp, lambda = lam)

11 out of 13 coefficients are nonzero
n = 203, number of events = 76
```

Again, we use the `fit2` object to plot the predicted individual and average survival curves (Figure 2.4).



(a) sorlie

(b) cgd

Figure 2.4: Predicted individual and average (red) survival probabilities for `sorlie`, left, and `cgd`, right, data sets.

```
R> surv_avg <- pcoxsurvfit(fit2)
R> surv_df <- with(surv_avg, data.frame(time, surv))
R> surv_ind <- pcoxsurvfit(fit2, newdata = cgd)
R> splot_cgd <- (plot(surv_ind, lsize = 0.05, lcol = "grey") +
+    geom_line(data = surv_df, aes(x = time, y = surv, group = 1),
+    col = "red") +
+    labs(caption = "(b) cgd") +
+    theme(plot.caption = element_text(hjust=0.5, size=rel(1.2)))
+ )
R> print(splot_cgd)
```

2.3.3 Simulated data set

In this section, we test our package on simulated data with time-dependent covariates, and compare its performance with that of the **penalized** algorithm. We first describe our data simulation process and then report the performance results.

We provide a user-friendly wrapper, `simtdc`, for the extended permutational algorithm for simulating time-dependent covariates provided by the **PermAlgo** package (Sylvestre and Abrahamowicz 2008).

We simulated a data set for 120 unique individuals with a follow-up time of up to 10 time units (years), 100 time-dependent and 900 time-fixed covariates — all drawn from a normal distribution with a mean of 0 and standard deviation of 1 with the true effect size, expressed on log hazard scale, of each covariate drawn from a uniform distribution [0, 2]. Since some individuals were observed in more than one time interval there were 444 observations, of which we used 299 observations for training and the remainder for testing. Covariates affected relative hazard only; event times were chosen assuming a constant total hazard rate of 0.2. Censoring times were chosen uniformly over the time period.

We compared our proximal gradient descent algorithm, **pcoxtime**, to the combination gradient descent-Newton-Raphson method from **penalized** (Goeman 2010). The two packages use different elastic net penalty specifications (**penalized** uses λ_1 and λ_2 for the lasso and ridge penalties instead of an overall λ and a mixing parameter α). We used $\alpha = 0.5$ and the default range of λ values in **pcoxtime**, then used a convenience function from [the development version of] **pcoxtime** to calculate values of λ_1 and λ_2 for **penalized**. Although the two approaches are similar, **penalized** has two possible limitations: (1) it cross-validates elastic net in two steps, finding a value of the ridge penalty λ_2 for each value of the lasso penalty λ_1 in order to fit an elastic net; for k-fold CV, this two-step procedure will require k times as much computational effort. (2) Possibly to compensate for this inefficiency, it uses Brent's algorithm to search for the optimal hyperparameter values (rather than using a parameter grid as we and others do), which risks converging to a local optimum (Goeman 2018).

To compare the predictions of **penalized** and **pcoxtime**, we used two approaches to choose the hyperparameters for **penalized**: (1) "**pcoxtime-**λ_1-based", using the optimal α and λ chosen by **pcoxtime** model to calculate the λ_1 and λ_2 values for the **penalized** model (we call this the *pcox-pen* model). (2) "**penalized-**λ_1-based", training the penalized model using the optimal λ_1 value determined by **penalized** from a vector of λ values generated from **pcoxtime**'s cross-validation (we call this the *pen-pen* model). The two predictions were then compared to **pcoxtime**'s estimates (*pcox*).

The two **pcoxtime** fits (*pcox-pen* and *pen-pen*) gave very similar estimates of the optimal λ (16.31 and 14.94, respectively). Comparing these results with **pcoxtime**'s, all three approaches gave similar estimates and confidence intervals for Harrell's C-statistic (Therneau 2020) (both *pcox* and *pcox-pen* gave $C = 0.65[0.51, 0.77]$, while the *pen-pen* values differed by a few percent: $C = 0.64[0.50, 0.74]$).

The **pcoxtime** package took about 30 times as long as **penalized** to compute the complete solution path (1162 seconds vs. 38 seconds), probably because our current implementation uses C++ only for likelihood computation and coefficient estimation for each λ; the solution paths are computed in R. All computations were carried out on an 1.80 GHz, 8 processors Intel Core i7 laptop. Table 2.1 compares the features of the different packages available for fitting penalized CPH models.

2.4 Comparison among packages

In this section, we compare the capabilities of **pcoxtime** to some of the most general and widely used R packages for penalized CPH models — **glmnet**, **fastcox** and **penalized**. Computational speed is important in high-dimensional data analysis; packages using coordinate descent based methods (**glmnet** and **fastcox**) are usually much faster than gradient descent based methods (**pcoxtime** and **penalized**) (Simon et al. 2011). Table 2.1 summarizes some important capabilities of the packages.

	pcoxtime	**glmnet**	**fastcox**	**penalized**
Supported models				
Time-dependent covariates	yes	no	no	yes
Penalty parameterization	λ, α	λ, α	λ, α	λ_1, λ_2
Post model predictions				
Survival and hazard functions	yes	no	no	yes
Model diagnostics & validation (prediction error, Brier score, calibration plots, etc.)	yes	no	no	no

Table 2.1: Capabilities of **pcoxtime**, **glmnet**, **fastcox** and **penalized** packages.

2.5 Conclusion

We have shown how the penalized CPH model can be extended to handle time-dependent covariates, using a proximal gradient descent algorithm. This paper provides a general overview of the **pcoxtime** package and serves as a starting point to further explore its capabilities.

In future, we plan to improve the functionality of **pcoxtime**. In particular, we plan to implement a coordinate descent algorithm in place of the current proximal gradient descent approach, which should greatly improve its speed.

Acknowledgments

We would like to thank Mr. Erik Drysdale and Mr. Brian Kiprop for the discussions and valuable feedback on earlier versions of the package. This work was supported by a grant to Jonathan Dushoff from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery.

Chapter 3

A comparison of machine learning methods to predict survival times for cancer patients: Incorporating time-varying covariates

The text I present here is a manuscript already submitted for publication

Author Contributions

SC performed statistical analyses with helpful feedback from JD, HS and BMB; SC wrote the first draft of the manuscript, and all authors revised the manuscript.

A comparison of machine learning methods to predict survival times for cancer patients: Incorporating time-varying covariates

Steve Cygu[1,*], Hsien Seow[2], Jonathan Dushoff[1,3], Benjamin M. Bolker[1,3]

[1]McMaster University, School of Computational Science and Engineering, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[2]McMaster University, Department of Oncology, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[3]McMaster University, Department of Biology, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[*]cygubicko@gmail.com

Abstract

The Cox proportional hazard model is commonly used in evaluating risk factors in cancer survival data. The model assumes an additive, linear relationship between the risk factors and the log hazard. However, this assumption may be too simplistic. Further, failure to take time-varying covariates into account, if present, may lower prediction accuracy. In this retrospective, population-based, prognostic study of data from patients diagnosed with cancer from 2008 to 2015 in Ontario, Canada, we applied machine learning-based time-to-event prediction methods and compared their predictive performance in two sets of analyses: 1) yearly-cohort-based time-invariant and 2) fully time-varying covariates analysis. Machine learning-based methods – gradient boosting model (gbm), random survival forest (rsf), elastic net (enet), lasso, ridge, and deepsurv neural network (nnet) – were compared to the traditional Cox proportional hazard (coxph) model and the prior study which used the yearly-cohort-based time-invariant analysis. Using Harrell's C index as our primary measure, we found that using both machine learning techniques and incorporating time-dependent covariates can improve predictive performance. Gradient boosting machine showed the best performance on test data in both time-invariant and time-varying covariates analysis.

3.1 Introduction

Early diagnosis and accurate prognosis can improve the clinical management of cancer patients. Good prognostic tools can help in treatment planning (Seow et al. 2011), aid communication with patients and patients' decision-making about surgery and treatments; and also help in timely and effective symptom management (Papachristou et al. 2018). Measuring cancer patients' wellbeing is significant in assessing response to treatment and capabilities for various types of care (Hayward et al. 2010). For instance, integrating palliative care interventions with oncological care for advanced cancer patients can lead to improved quality of life, reduced symptom burden, fewer hospital visits, and reduced health costs (Seow et al. 2011; Seow et al. 2016; Seow et al. 2020). Predictive computational methods that predict symptoms, patients' wellbeing, and survival time can help clinicians customize treatment regimes and give timely interventions.

Traditional statistical methods such as Kaplan-Meier and Cox proportional hazard (CPH) models have been used to model survival data (Cox 1972; Fujino et al. 2003; Seow et al. 2020). These techniques estimate the probabilities of survival by assuming an additive, linear relationship between the risk factors (covariates) and the log hazard, i.e., proportional hazard assumption. In clinical survival data, three specific challenges have been identified (Simon et al. 2011; Ishwaran et al. 2014; Montazeri et al. 2016): 1) high dimensional data (large number of features), 2) data censoring (i.e., time-to-event is imperfectly observed), and 3) violation of the proportional hazard assumption. Studies on the accuracy of clinical prediction of survival time have found poor agreement with the actual survival times, showing that practitioners' predictions tend to be longer than actual survival times (Chow et al. 2001; Cheon et al. 2016; Seow et al. 2020).

Additional challenges in accurate prediction of survival of cancer patients emerge from the growing complexity of cancer, variant treatment options, heterogeneous patient populations and failure to account for measurements which change over time (time-varying covariates) (Seow et al. 2020). In survival data, time-varying covariates are common. For example, cancer patients' chemo-therapy treatment plan or healthcare access may change over the course of the study. The standard

CPH model assumes that the covariates are time-invariant and have a constant linear effect over the entire follow-up period (Cygu et al. 2021; Yao et al. 2021). Time-invariant CPH models have been extended to handle time-varying covariates (Andersen and Gill 1982).

The use of ML methods in predicting the risk of death of cancer patients from clinical data is not new (Gupta et al. 2014; Kourou et al. 2015; Montazeri et al. 2016; Mihaylov et al. 2019). Depending on how these methods are applied, they can be considered standard ML methods (directly applied to predict the outcome of interest such as survival status) or hazard-based ML methods (modified to handle time-to-event data). Standard ML methods use binary classification to predict the survival status of subjects within a particular time window. Since binary classifiers consider only whether or not the event occurred in the last observation window, they lack the interpretability and flexibility of models that consider hazards as a function of time (Cygu et al. 2021). Most hazard-based ML methods, such as artificial neural networks (ANN) (Katzman et al. 2018b), survival trees and random forest (Wang et al. 2019; Bou-Hamad et al. 2011), predict events of interest using covariates measured at the time of diagnosis, not accounting for the time-varying covariates. Furthermore, only a few of these models have incorporated patient-reported cancer diagnosis related outcomes such as level of pain as potential covariates to build predictive models (Seow et al. 2020).

In this paper, we build, validate and compare traditional CPH models and hazard-based ML models, using both time-invariant and time-varying covariates (including both clinical and patient reported variables). We compare the performance of our models to the prior yearly-cohort-based time-invariant traditional CPH model with backward variable selection implemented by Seow et al. (Seow et al. 2020).

3.2 Results

3.2.1 Performance of the machine learning models

The results of our comparisons on training and testing datasets are shown in Figure 3.1, summarized by the 2.5%, 50% and 97.5% quantiles of the estimated Harrell's C index on 200 bootstrap resamples of the respective datasets. The details of these comparisons are given in the Model evaluation and comparison section. The ML algorithms we used fall into 4 groups: penalized Cox model, i.e., elastic net (enet), lasso and ridge; gradient boosting machine (gbm); neural network (nnet); and random survival forest (rsf). We also compare with the traditional (non-ML) CPH model used by Seow et al. (Seow et al. 2020), i.e., traditional CPH model with backward variable selection (BS coxph) as well as full traditional CPH (full coxph). Due to computational and implementation constraints, rsf and nnet were not implemented on the fully time-varying covariates analysis.

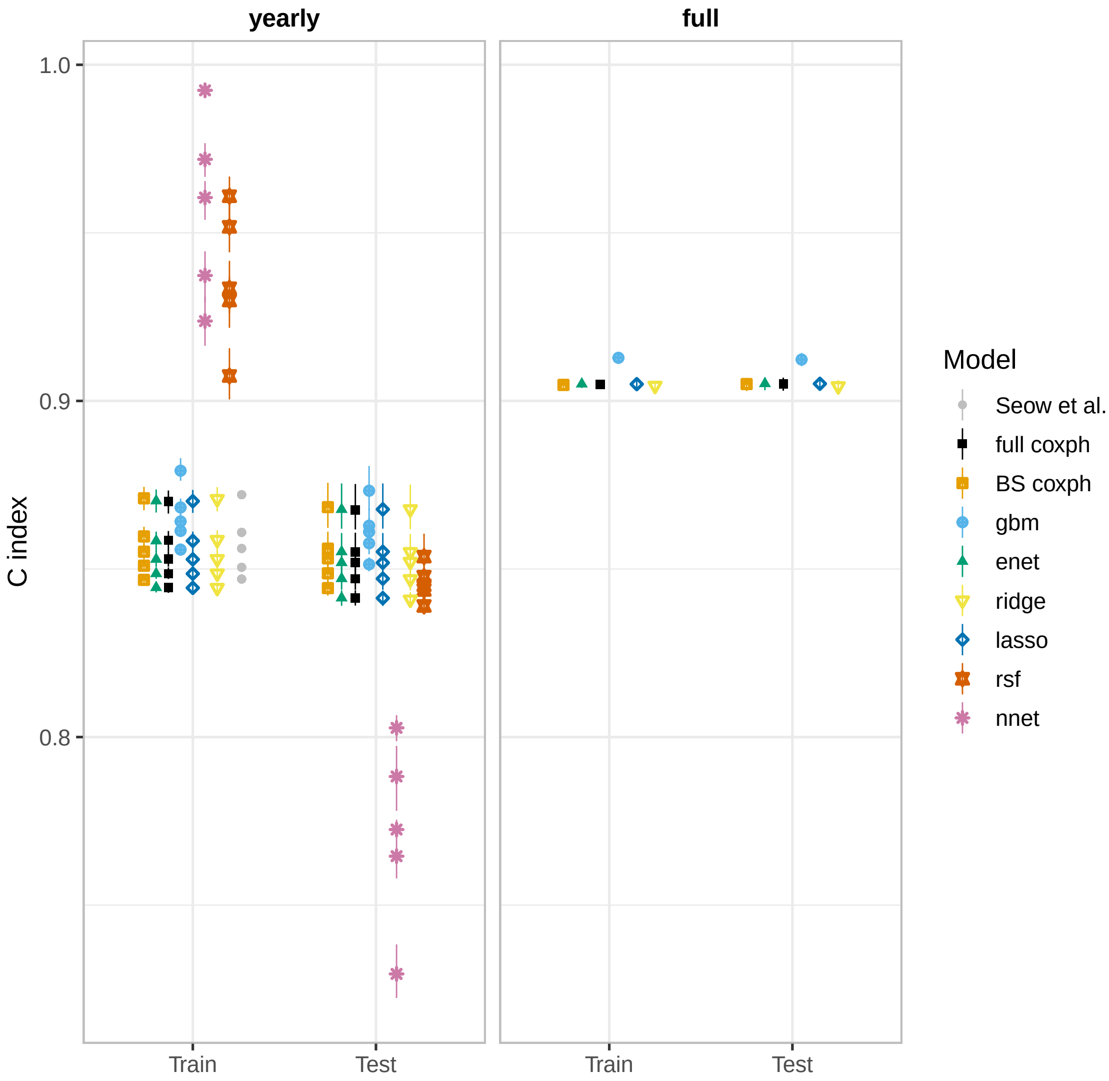


Figure 3.1: A comparison of Harrell's concordance scores (C index) for the yearly-cohort-based time-invariant (yearly) and fully time-varying (full) covariates analysis. Higher values are better. The yearly-cohort-based estimates obtained in Seow et al. (Seow et al. 2020) were available for training data only. For comparison, we provide both training and test C index scores. Generally, gradient boosting machine slightly performs better than all the other models. On the other hand, random survival forest and neural network models overfit.

On the yearly-cohort-based time-invariant analysis, gradient boosting machine (gbm) had the highest score in the test data across all the cohorts. For the training data, the BS coxph method matches the earlier results almost exactly (as expected). Neural network (nnet) had the highest score on the training data but performed the worst on the testing data, indicating that the fit was not reliable.

On the fully time-varying covariates analysis, gbm's predictive performance was again higher than all other models, on both training and testing data. The performance of the other models (including the traditional models) was similar. The models were generally able to achieve better prediction when using the fully time-varying covariates than when using yearly-cohort-based time-invariant covariates. Separate cohort comparisons are provided in Supplementary Figure S1.

3.2.2 Temporal performance of the models

To evaluate the performance of the models at different survival marks, we use the time-dependent AUC. Figure 3.2 shows the distributional summary of the time-dependent AUC achieved in 200 replicates of bootstrapped samples of the test data. The central point represents the median (50%) quantile, while the lower and upper ends of the lines represent lower (2.5%) and upper (97.5%) quantiles of the estimates. Higher values indicate better performance, narrower ranges indicate more stable algorithms.

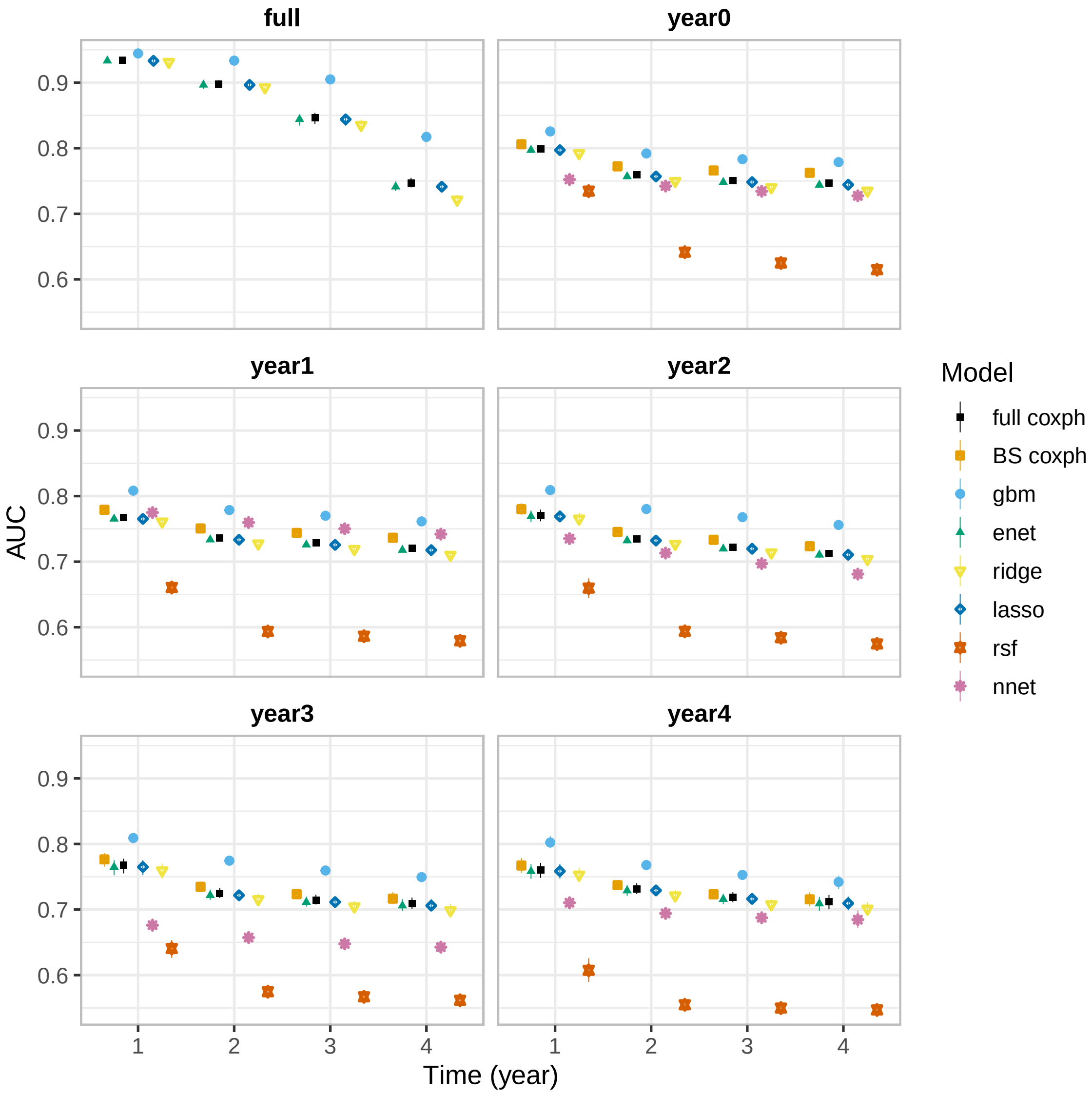


FIGURE 3.2: Time-dependent AUC scores evaluated at different time points. The scores are based on 200 bootstrapped samples (50 for nnet and rsf due to computational limitations) of the test data. Models with higher scores and narrower confidence intervals are better performers. As with the concordance index, models do better with the fully time-varying covariates (full), and gbm does better than other models.

In both fully time-varying and yearly-cohort-based time-invariant covariates, gbm model has slightly better estimates with comparatively narrower confidence intervals. Estimates based on the fully time-varying covariates are generally better than estimates based on the yearly-cohort-based time-invariant covariates.

3.2.3 Most prognostic predictors

Figure 3.3 shows permutation-based importance scores (see Methods) for the top 15 features in each data set for both gbm (the top performing model), and for BS coxph (for the benchmark model). Palliative care, cancer type, age and cancer stage were identified as important in all cases.

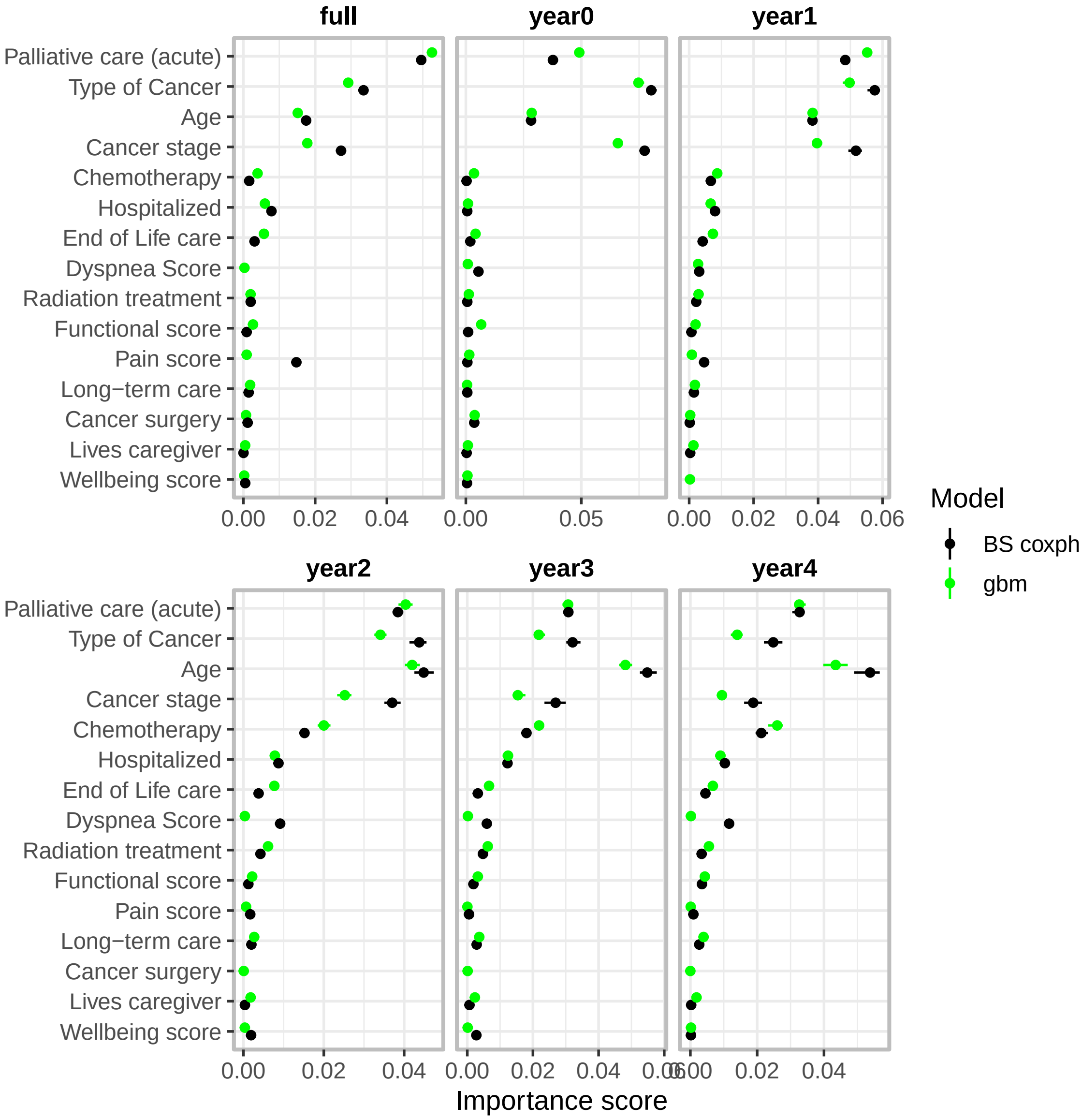


FIGURE 3.3: Variable importance scores together with the corresponding 2.5%, 50% and 97.5% quantiles, based on the best (gbm) and benchmarking models (BS coxph). Notably, palliative care, age of the patient, cancer type and cancer stage stood out, across the cohorts, as some of the most important prognostic factors on survival of cancer patients.

To compare the features across the cohorts, we ranked all the features and counted the number of times each feature was among the top 5 across all the models and cohorts (Figure 3.4).

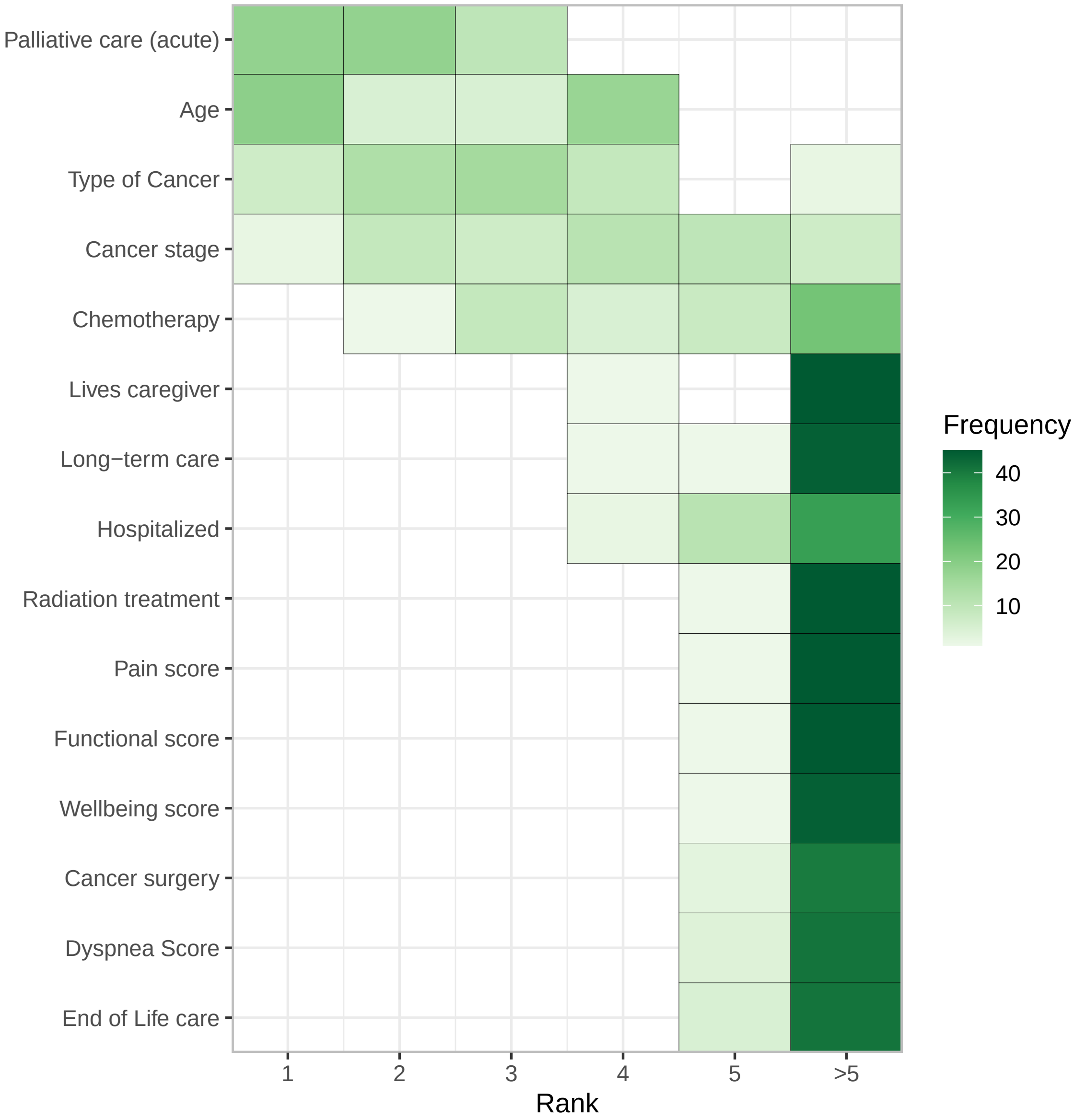


Figure 3.4: The number of times, frequency, a given feature is ranked, on top 5, by a particular model in a given cohort as one of the most important feature. Low rank means a particular feature is predictive and hence important.

3.3 Discussion

Compared to machine learning algorithms, traditional CPH models are less suited for prediction, although the full (unselected) CPH may be better for inference about the impact of a specific predictor. If our interest is to predict time-to-event of cancer patients based on a number of clinical, medical and self-reported predictors, machine learning-based models may be preferable over traditional CPH models in analyses that consider more data at once.

Fitting models which incorporate fully time-varying covariates require special attention; only a subset of the models fitted in the yearly-cohort-based time-invariant models could support this kind of analysis. Thus, in the full dataset incorporating time-varying covariates, in addition to traditional CPH, only gradient boosting and penalized models were implemented.

Harrell's C index was measured on mutually exclusive training and testing datasets. For overfitted models, we would expect the model to perform well on the training set but poorly on the testing set. Other than random survival forest and deepsurv neural network, none of the models appeared overfitted. Penalized models did not show major improvement in predictive performance over the traditional CPH model with backward variable selection model which was slightly better than the full traditional CPH model.

Our results show that ML-based methods can provide more accurate alternatives to traditional hazard-based methods in both yearly-cohort-based time-invariant and full time-varying covariates. However, CPH model with backward variable selection performed comparably to the ML-based methods, as has also been seen elsewhere (Spooner et al. 2020; Seow et al. 2020). Cox model with gradient boosting machine had the highest predictive performance score in all the comparisons done. This model has additional advantages of computational efficiency and fewer hyperparameters when compared to methods like random forests and neural networks.

In summary, time-varying covariates greatly improve model prediction, and not only in the ML context. We also find that gradient-boosting machine (gbm) improves performance across both the cohort and time-varying approaches, suggesting that it may be a good choice in general for problems of this nature.

3.4 Methods

3.4.1 Study participants

Subjects were adults diagnosed with cancer from a population-based, retrospective prognostic study, as confirmed by the provincial cancer registry in Ontario, Canada, from January 1, 2008, to December 31, 2015.

The study was reviewed by Hamilton Integrated Research Ethics Board and deemed exempt because it used de-identified secondary data.

Patients and the public were not involved in this research. It used de-identified, secondary administrative data analysis, (which is allowed to be used for research purposes), and thus patient consent was not obtained. Seow et al. (Seow et al. 2020) provide a detailed description of the data and study setting.

3.4.2 Availability of data and materials

The de-identified administrative data are not publicly available and may be obtained from a third party, ICES (formerly the Institute for Clinical Evaluative Sciences) for researchers who meet the criteria for permissible access. These data represent secondary data analysis and are not owned or collected by the study authors. A data request can be sent here: https://www.ices.on.ca/About-ICES/ICES-Contacts-and-Locations/contact-form.

We provide all the R codes used for the analysis in the form of a workflow R package for the analysis of similar datasets, which can be accessed on GitHub through https://github.com/CYGUBICKO/satpred.

3.4.3 Data pre-processing

A number of pre-processing steps were undertaken to prepare the dataset for modelling. To avoid excluding cases or variables from the dataset, a “missing” category was created for the patient-reported categorical variables. Numerical variables, such as age, were mean-centered.

3.4.4 Analysis plan

We performed two classes of analysis which were based on the structure of the data, *yearly cohort* (yearly-cohort-based covariates) and *full dataset* (fully time-varying covariates), as summarized in Figure 3.5.

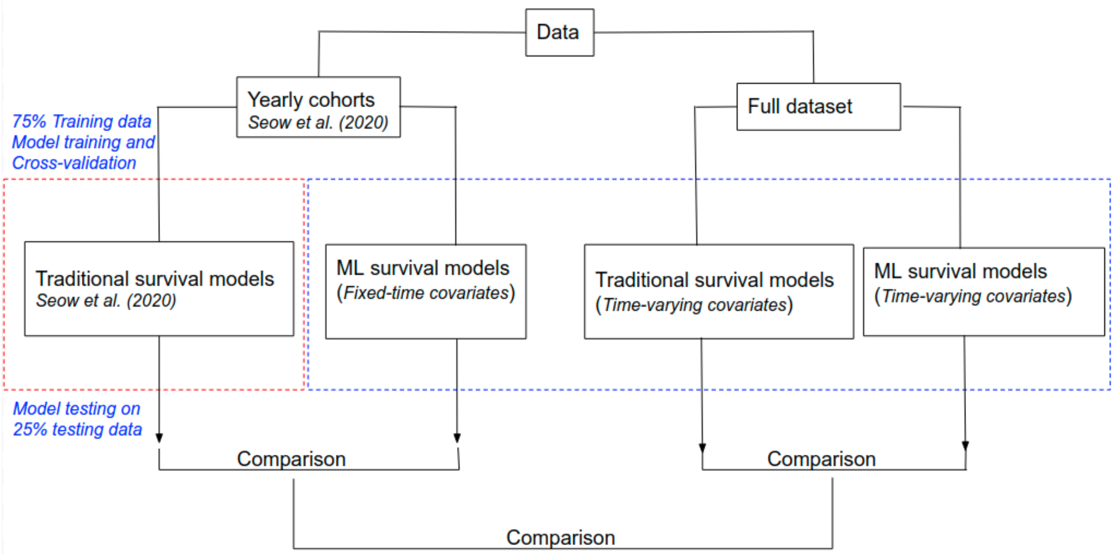


Figure 3.5: The blue dotted rectangle indicates the major analytical contribution of this paper. We also replicated analyses by Seow et al. (Seow et al. 2020), indicated by red rectangle. We performed two classes of analysis depending on the nature of the dataset. The first set of analyses closely followed modelling procedure in Seow et al. (Seow et al. 2020) which is based on yearly cohorts (we refer to these as yearly cohort models) and, by construction, takes care of changing covariates over the observation period. We then used hazard-based ML models and compared predictive performance to prior results in Seow et al. (Seow et al. 2020), which used traditional CPH model with backward variable selection. The second set of analyses used both traditional CPH and hazard-based ML models which directly incorporate time-varying covariates on the full dataset. We refer to these as full dataset models. We also compared predictive performance of the full dataset models to those of yearly cohort models.

Yearly cohort models

To account for changing covariates in a traditional CPH model, Seow et al. (Seow et al. 2020) created yearly cohort datasets. Each year that a patient survived post-diagnosis, they were entered into a separate cohort model; thus, only patients who survived to a certain point (survival mark) contributed to the corresponding conditional analysis. Cohort definitions are summarized in the Figure 3.6. Our first set of hazard-based ML models used these yearly cohort datasets and compared the predictive performance with those obtained from traditional CPH models fitted in the prior analyses by Seow et al. (Seow et al. 2020). In addition to our ML fits, we also replicated the traditional CPH model together with the backward variable selection procedure in Seow et al. (Seow et al. 2020) with slight modifications. For instance, our models used 75% − 25% as opposed to 60% − 40% *train–test* partition.

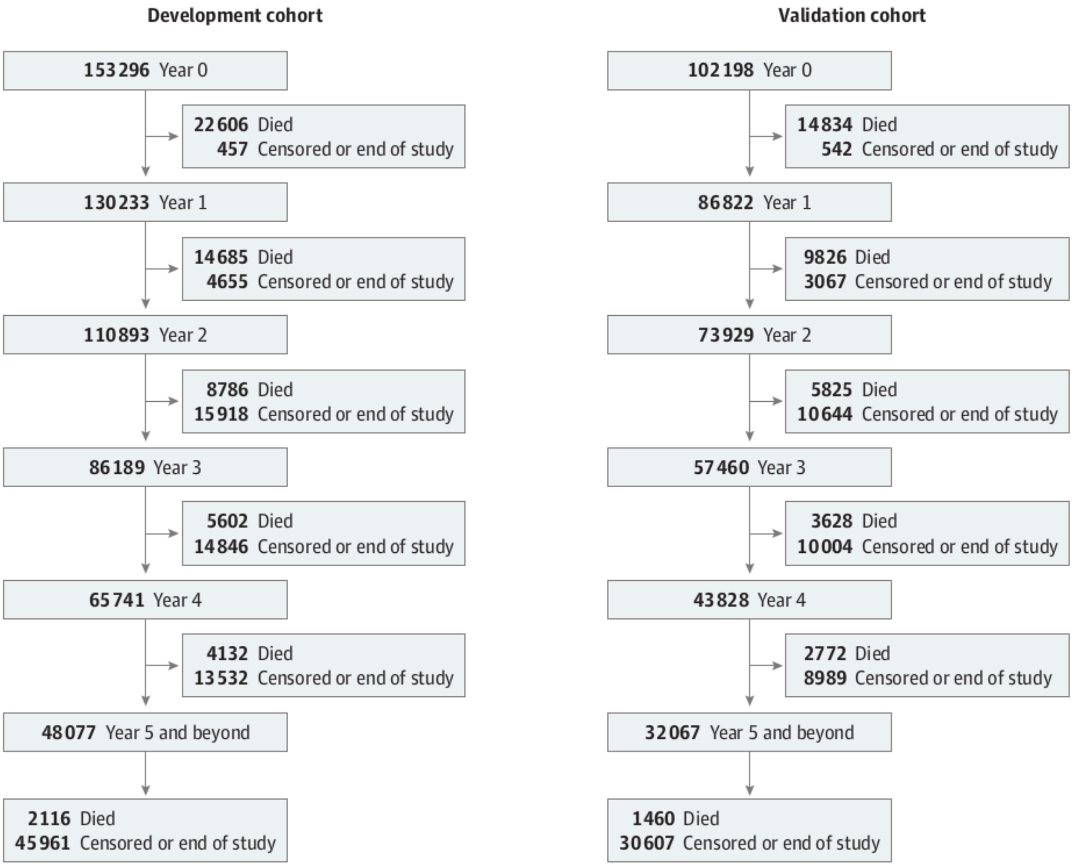

Development cohort
Validation cohort
153296 Year 0
22606 Died
457 Censored or end of study
130233 Year 1
14685 Died
4655 Censored or end of study
110893 Year 2
8786 Died
15918 Censored or end of study
86189 Year 3
5602 Died
14846 Censored or end of study
65741 Year 4
4132 Died
13532 Censored or end of study
48077 Year 5 and beyond
2116 Died
45961 Censored or end of study
102198 Year 0
14834 Died
542 Censored or end of study
86822 Year 1
9826 Died
3067 Censored or end of study
73929 Year 2
5825 Died
10644 Censored or end of study
57460 Year 3
3628 Died
10004 Censored or end of study
43828 Year 4
2772 Died
8989 Censored or end of study
32067 Year 5 and beyond
1460 Died
30607 Censored or end of study

Figure 3.6: The yearly cohorts defined by Seow et al. (Seow et al. 2020).

Full dataset models

Both traditional CPH and hazard-based ML models which incorporate time-varying covariates require the dataset to be in a specific format – that is, counting process format (Thomas and Reyes 2014; Allison 2010; Fox 2002). The data are expanded from one record-per-subject to one record-per-interval between each event time, per subject, such that each record corresponds to the interval (365 days in this case) of time during which the entries of time-varying covariates are treated as constant. Once this special dataset, which combines all the yearly cohorts, has been constructed, the event time is now defined by $(\text{start}, \text{stop}]$ interval during which the subject was continuously at risk of the event, and events can only occur at the end of the interval. For example, for an individual who survived for 3 years, the "at-risk" interval is defined as $(0, 365], (365, 730]$ and $(730, 1095]$, representing the segments in which they are event free and uncensored.

Currently, only penalized, gradient boosting machine and random forest implementations support time-varying covariates for survival analysis. Due to computational challenges, we only fitted and compared penalized, i.e., lasso, ridge and elastic net and gradient boosting machine models, in addition to the traditional CPH model. All computations were carried out on a server with 4 clusters, each with 8 Intel Xeon 3.40 GHz CPUs and a 128 GB RAM.

3.4.5 Model selection

A total of 4 classes of machine learning algorithms and traditional Cox proportional hazard models capable of handling censored data were used in this analysis. The outcome of interest was time to death (days) as recorded in the Vital Statistics database (Seow et al. 2020). The following classes of models were trained and evaluated:

1. Traditional Cox proportional hazard model: Implemented for both time-invariant and time-varying covariates.

2. Penalized Cox regression: lasso, ridge and elastic net. Implemented for both time-invariant and time-varying covariates.

3. Random survival forests: Capable of handling both time-invariant and time-varying covariates but requires large amount of computer memory for large datasets due to large forests constructed during model training. Due to this

challenge, we trained random forest on only a subset of data (2000 cases) for each cohort in the time-invariant covariates analysis.

4. Generalized boosted regression models: Cox-based gradient boosting machine (gbm) were implemented for both time-invariant and time-varying covariates datasets.

5. Neural network: A multi-layer feed-forward neural network for survival data. Similar to random forest, this model also runs into memory issues and consequently, was trained on a subset of data and, currently, capable of handling only time-invariant covariates.

A brief description of these algorithms can be found in the Supplementary Methods S1.

Each of the hazard-based ML algorithms outlined above has at least one hyper-parameter and, as result, requires parameter tuning. For this, we perform 10-fold cross validation. For penalized approaches (lasso, ridge and elastic net), the hyper-parameters are tuned using cross-validated partial log-likelihood; for random survival forest, neural networks and gradient boosting machine, Harrell's concordance index C is used. For the Cox proportional hazard model, we apply stepwise variable elimination on the multivariate model which fits all the covariates and identifies a subset of important variables according to Akaike's information criterion. The final CPH model is then fitted using only those variables selected in the stepwise procedure. A list of hyper-parameters that were tuned can be found in Supplementary Table S1, together with the R packages used to implement each of the models.

3.4.6 Model evaluation and comparison

The models implemented in this work have different strengths and limitations in terms of assumptions, interpretability, computational efficiency, etc. In this work, we focus on comparing the predictive accuracies of these models. We implemented the following metrics to evaluate and compare the performance of our models on the test data:

- **Harrell's concordance index (C index)**. In survival analysis, a pair of patients is called concordant if the risk of the event predicted by a model is lower for the patient who experiences the event at a later time-point. The concordance index is the frequency of concordant pairs among all comparable pairs of subjects. Pairs are incomparable if their event times are equal, or if either subject is censored before the other subject experiences an event (Therneau 2022). Harrell's C index can be used to measure and compare the discriminative power of a risk prediction models (Harrell Jr et al. 1996). It provides a holistic measure of the model performance over the entire time period, while accounting for censoring.

- **Time-dependent AUC**. The Receiver Operating Characteristic (ROC) curve and the associated area under curve (AUC) are widely used in medical research to quantify the discriminating power of machine learning models. The ROC curve plots the probability of both true positive (proportion of positive class correctly classified by the model) and the false positives (proportion of the negative class incorrectly classified by the model) at various cut off values of the risk score. The AUC summarizes the probabilities of true and false positives over all possible cut off values into a value ranging between 0 and 1; and gives an overall measure of predictive accuracy of a predictive model. The standard ROC considers the event status and the risk scores as fixed over time; however, in many medical applications, these quantities may change over the follow-up time; in such situations, binary classification of cases (as true positive and true negative) without taking into account the time-to-event may be inappropriate. Heagerty et al. (Heagerty et al. 2000) proposed a time-dependent ROC which extends the standard ROC curve analysis for binary outcome data to time-to-event data (see Supplementary Methods S2).

To evaluate the sensitivity and uncertainty of the predictive performance measures, we applied bootstrap resampling to estimate the 2.5%, 50% and 97.5% quantiles of the distribution of the scores. We used 200 bootstrap resamples of both training and test datasets. The training estimates were included for comparison with those reported in Seow et al. (Seow et al. 2020).

3.4.7 Identifying prognostic features

A permutation-based variable importance score was used to identify the most important prognostic features. For each replicate, we randomly resample the values of a focal predictor and record how our metric changes due to this perturbation. The key idea is that if a particular predictor has high power to predict the response, then randomly permuting its observed values will lead to a considerable change in the predictive accuracy of the model. In this case, we conclude that this predictor is important. In our implementation, Harrell's C index is used as a measure of predictive power.

Acknowledgements

This work was supported by a grant to Jonathan Dushoff from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery. The study was partially supported by the Canadian Institutes for Health Research (Grant #379009 and #383402), and also supported by ICES, formerly known as the Institute for Clinical Evaluative Sciences, which receives funding from the Ontario Ministry of Health and Long Term Care. It also uses data from Cancer Care Ontario, Canadian Institutes of Health Information, IQVIA Solutions Canada Inc, and the Ontario Association of Community Care Access Centers. The opinions, results, and conclusions reported in this paper are those of the authors solely and the funders and data sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author contributions statement

S.C. designed and implemented machine learning experiments and wrote the first draft of the paper. H.S. provided the data, expert guidance and reviewed the manuscript. J.D. provided expert guidance and critical review of the manuscript.

B.M.B supervised the implementation of machine learning experiments and provided expert guidance for all aspects. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at . . . .

Competing Interests: The authors declare no competing interests.

A comparison of machine learning methods to predict survival times for cancer patients: Incorporating time-varying covariates

Steve Cygu[1,*], Hsien Seow[2], Jonathan Dushoff[1,3], Benjamin M. Bolker[1,3]

[1]McMaster University, School of Computational Science and Engineering, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[2]McMaster University, Department of Oncology, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[3]McMaster University, Department of Biology, Hamilton, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

[*]cygubicko@gmail.com

3.5 Supplementary Methods S1: Machine learning algorithms

Standard machine learning methods use binary classification to predict the status of a subject. Such binary classifiers lack the interpretability and flexibility of models that consider event times explicitly (Cygu et al. 2021). Several machine-learning algorithms have been extended to estimate hazards and use a survival-model paradigm to consider event times (and handle observation censoring). Below, we summarize the traditional Cox proportional hazard model approach and give a short description of some hazard-based ML approaches.

3.5.1 Cox proportional hazard model (coxph)

Time-invariant covariate survival data is often presented in the form $\{t_i, \delta_i, x_i\}_{i=1}^{n}$, where t_i is the observed event time for individual i, δ_i is an indicator variable for censored or observed event of interest, and x_i is a vector of covariates. Traditional hazard-based methods such as Cox proportional hazard (CPH) model (Cox 1972) is commonly used in survival data. The CPH model defines the hazard function at time t as

$$h_i(t) = h_0(t) \exp(x_i^\top \beta), \tag{3.1}$$

where $h_0(t)$ is the non-parametric baseline hazard function and $\exp(x_i^T \beta)$ is the relative hazard, which summarizes the effects of the covariates. Under the proportional hazard assumption, the model parameters can be estimated by fitted by minimizing a partial log-likelihood which does not involve the non-parametric baseline hazard.

When covariates change over time during the follow-up period, the observed survival data is of the form $\{t_i^{\text{start}}, t_i^{\text{stop}}, \delta_i, x_i(t)\}_{i=1}^n$. The only difference is that x_i is now a (piecewise constant) function of time, and Equation 3.1 adjusted appropriately for downstream implementation (Cygu et al. 2021; Harrell Jr 2015).

Although CPH models are commonly used, especially when the main goal is to make inference on how the covariates impact on the survival probabilities, they assume a linear relation between the risk factors (covariates) and event hazard. This assumption may be too simplistic in some contexts.

3.5.2 Penalized cox proportional hazard models

Traditional CPH models may overfit, particularly in the case of high-dimensional data. Penalized methods such as **lasso**, **ridge** and **elastic net** offer a convenient way of addressing overfitting. Lasso and elastic net can also be used to select a subset of useful predictive feature while eliminating others.

Penalized methods add a penalty to the log-likelihood function, which has the effect of shrinking the coefficient values towards zero, reducing sampling variance and reducing the impact of less important features on the model. The ℓ_1 (lasso) penalty is based on the absolute value of the coefficients. It typically reduces the number of predictors used (by assigning zero coefficients to a subset of predictors). The number of features selected by lasso is bounded by number of observations (Tibshirani 2013). On the other hand, ℓ_2 (ridge) penalty is based on the square of the coefficients. The ridge penalty shrinks the coefficients towards (but never all the way to) zero. The elastic net (a combination of ℓ_1 and ℓ_2) penalties combines the strength of lasso and ridge for improved predictive performance (Simon et al. 2011).

Lasso, ridge and elastic net regression have all been extended to handle time-varying covariates survival data and are evaluated here. See Cygu et al. (Cygu et al. 2021) for a detailed discussion on penalized Cox proportional hazard model for time-dependent covariates.

3.5.3 Random survival forest (rsf)

Random forests are ensembles of decision trees that are grown on bootstrapped training samples of the original data by choosing m random samples of the original set of p predictors at each split (node). Random survival forests (RSF) are random forests adapted for survival analysis of censored data. In a random survival forest, the feature and split point chosen is the one that maximizes the survival difference between daughter nodes i.e., that maximizes the log rank statistic over all available split points and features (Ishwaran et al. 2008). As opposed to CPH-based approaches, which assume linear combination of the covariates, RSF are capable of automatically handling and identifying non-linear and complex interactions. RSF are free of assumptions and due to the randomization during splitting, and lends itself to feature selection through measures of variable importance. A possible drawback of RSF is the bias in splitting in the presence of predictors with multiple possible split points, e.g., categorical predictors with many levels (Spooner et al. 2020). In our implementation, we ran into computational memory issues due to large forests constructed in model training. As a result, we trained RSF on only a subset of the time-invariant cohorts dataset. A detailed description of RSF is outlined in Ishwaran et al. (Ishwaran et al. 2008). We used the following hyperparameters for our random survival forest fits:

- The number of trees to grow, **ntree**.
- The number of variables randomly selected for splitting a node, **mtry**.
- Minimum size of a node (after splitting), **nodesize**.
- The rule for splitting nodes, **splitrule** (log rank or log rank score).

Yao et al. (Yao et al. 2021) generalized the conditional inference and relative risk survival forests to incorporate time-varying covariates and proposed a more general framework for estimating survival function in the presence of time-varying covariates. However, due to computational limitations, we did not implement this method.

3.5.4 Generalized boosted regression models (gbm)

Boosting is an iterative method which uses an ensemble technique to train weak learners sequentially, where each new model that is added to the ensemble learns from the "mistakes" of the previous models.

There are two main approaches to boosting in survival analysis: likelihood-based and gradient boosting. Likelihood-based boosting uses base learners that maximize the overall likelihood in each boosting step, selecting only the base-learner which leads to largest increase in the likelihood. On the other hand, gradient boosting is equivalent to iteratively re-fitting the residuals of the ensemble model at each step. With correct choice of boosting steps, boosted models are resistant to overfitting and work well in high-dimensional data (Spooner et al. 2020). In this work, we used gradient boosting machine with the following hyperparameters:

- The number of trees, **n.trees**.
- The shrinkage parameter, **shrinkage**, which controls the rate at which boosting learns. Small values of shrinkage require using large values of n.trees.
- The interaction depth, **interaction.depth**, which controls the complexity of the boosted ensemble i.e., the highest level of variable level interaction. A value of 1 implies an additive model, a value of 2 implies a model with up to 2-way interactions, etc.

3.5.5 Neural networks (nnet)

We implemented a multi-layer feed-forward neural network, **DeepSurv** (Katzman et al. 2018b), of which the output is the negative partial log-likelihood, parameterized by the weights of the networks. The hidden layers are fully connected, not necessarily of the same size, and are passed through nonlinear activation functions. The output layer has a single node with a linear activation which gives the output $\hat{h}_i(t)$ (log-risk hazard estimate). Due to computational limitations, we implemented this model on a subset of the data. In addition, the current implementation does not support time-varying covariates. We considered the following hyperparameters for tuning:

- The number of hidden layers, **layers**.
- Size of the hidden the layers, **num_nodes**, which defines the number of network weights and consequently the complexity.
- The dropout rate, **dropout**. Overfitting is a potential problem in neural networks. In particular, when there are multiple hidden layers, the problem is more often because of the large number of weights in comparison to the number of samples. Dropout is a technique that can be used to deal with this problem by randomly dropping some of the nodes, together with their connections (Srivastava et al. 2014).
- Stochastic gradient descent learning rate, **learning_rate**, which determines the step size of the weight iteration.

3.6 Supplementary Methods S2: Model evaluation methods

3.6.1 Time-dependent AUC

Let R be the estimated or predicted risk score, T denote the time to the occurrence of the event of interest and t define some time horizon. The individual's event status at t is defined as $D(t) = 1\{T \leq t\}$, which equals 1 if the event has occurred and 0 otherwise. Assuming that a higher value of R is associated with higher risk of event occurrence and that individual is predicted to have experienced event in the interval $(0, t]$ if $R > c$, where c is some cut off value, otherwise the individual is predicted to be event free in the interval $(0, t]$. Heagerty et al. (Heagerty et al. 2000) defined the sensitivity and specificity at the time horizon t as

$$\begin{aligned} \text{Sensitivity}(c, t) &= P(R > c | T \leq t) \\ \text{Specificity}(c, t) &= P(R \leq c | T > t). \end{aligned} \tag{3.2}$$

The major difference between the definition of sensitivity and specificity in standard binary case and the definition in Equation 3.2 is that the latter is defined with respect to the time horizon t. Sensitivity(c, t) and Specificity(c, t) are referred to as time-dependent sensitivity and specificity, respectively, and the resulting ROC curve (plot of Sensitivity(c, t) against $1 -$ Specificity(c, t)) is the time-dependent ROC curve at time horizon t. For our analysis, we used **R** software package **risksetROC** which implements this extension (Heagerty and Zheng 2005; Heagerty and Paramita Saha-Chaudhuri 2012).

3.7 Supplementary Figure S1: Harrell's C-index for each cohort

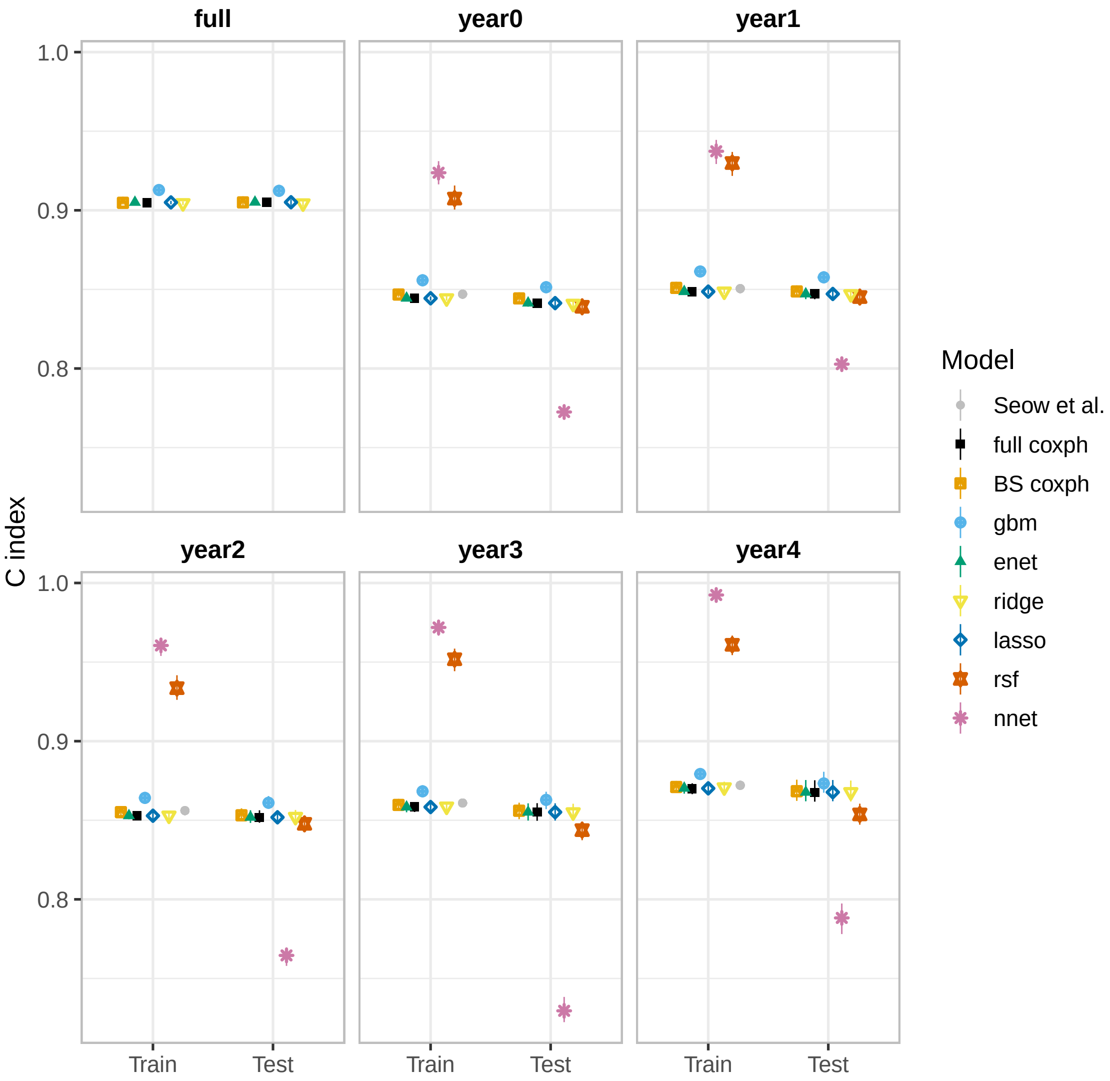


FIGURE 3.7: A comparison of Harrell's concordance scores (C index) for the full dataset and the yearly cohorts. Higher values are better. For comparison with Seow et al. (Seow et al. 2020), we provide both training and test C index scores. Generally, gradient boosting machine slightly performs better than all the other models. On the other hand, random survival forest and neural network models seem to overfit.

3.8 Supplementary Table S1: Tuning parameters

Models	**R package**	**Data**	**Tuned hyper-parameters**
Cox PH model; backward selection	survival; rms	All	
Ridge	glmnet	Year 0	alpha=0, lambda=0.0250
		Year 1	alpha=0, lambda=0.0238
		Year 2	alpha=0, lambda=0.0184
		Year 3	alpha=0, lambda=0.0162
		Year 4	alpha=0, lambda=0.0147
		Full	alpha=0, lambda=0.0203
Elastic Net	glmnet	Year 0	alpha=0.4, lambda=0.0024
		Year 1	alpha=0.8, lambda=0.0011
		Year 2	alpha=0.6, lambda=0.0011
		Year 3	alpha=0.8, lambda=0.0006
		Year 4	alpha=0.2, lambda=0.0014
		Full	Alpha=0.5, lambda=0.0007
LASSO	glmnet	Year 0	alpha=1, lambda=0.0012
		Year 1	alpha=1, lambda=0.0010
		Year 2	alpha=1, lambda=0.0007
		Year 3	alpha=1, lambda=0.0005
		Year 4	alpha=1, lambda=0.0004
		Full	alpha=1, lambda=0.0004
Random survival forest	randomForestSRC	Year 0	ntree=1000, mtry=20, nodesize=10, splitrule="logrank"
		Year 1	ntree=1500, mtry=20, nodesize=10, splitrule="logrank"
		Year 2	ntree=1500, mtry=20, nodesize=10, splitrule="logrank"
		Year 3	ntree=1000, mtry=20, nodesize=10, splitrule="logrank"
		Year 4	ntree=1000, mtry=20, nodesize=10, splitrule="logrank"
Gradient boosting machine	gbm and gbm3	Year 0	shrinkage=0.1, n.trees=3548, interaction.depth=2
		Year 1	shrinkage=0.1, n.trees=3541, interaction.depth=2
		Year 2	shrinkage=0.1, n.trees=2711, interaction.depth=2
		Year 3	shrinkage=0.1, n.trees=2112, interaction.depth=2
		Year 4	shrinkage=0.1, n.trees=1443, interaction.depth=2
		Full	Shrinkage=0.1, n.trees=2982, interaction.depth=2
Neural network	survivalmodels	Year 0	num_nodes=32, learning_rate=0.001, dropout=0.2
		Year 1	Layers=1, num_nodes=128, learning_rate=0.001, dropout=0.2
		Year 2	layers=1, num_nodes=128, learning_rate=0.01, dropout=0.2
		Year 3	layers=4, num_nodes=32, learning_rate=0.01, dropout=0.2
		Year 4	layers=1, num_nodes=128, learning_rate=0.01, dropout=0.1

Table 3.1: Tuning parameters

Chapter 4

Outcome plots: uncertainty estimation and bias correction for predictions and effects in simple and generalized linear (mixed) models

The text I present here is a draft of a manuscript planned for submission for publication

Author Contributions

SC performed statistical analyses with helpful feedback from JD and BMB; SC wrote the first draft of the manuscript, and all authors revised the manuscript.

Outcome plots: uncertainty estimation and bias correction for predictions and effects in simple and generalized linear (mixed) models

Steve Cygu[1, *], Benjamin M. Bolker[1,2], Jonathan Dushoff[1,2]

[1]School of Computational Science and Engineering, McMaster University, Hamilton, Ontario, Canada
[2]Department of Biology, McMaster University, Hamilton, Ontario, Canada

* cygubicko@gmail.com

Abstract

In generalized linear (mixed) models that involve complex multiplicative interactions, multi-parameter variables, additional non-focal predictors or a non-linear link function, outcome plots (prediction and effect plots) can aid in understanding difficult-to-interpret coefficient estimates. Outcome plots depend on the choices we make about the non-focal predictors. The most common approach is to generate estimates (central estimates, predictions and effects) at a reference point, usually the mean of non-focal predictor. We call this mean-based approach and it estimates effect of an average case in the population. In the presence of additional non-focal predictors, non-linear link functions, random effect terms, etc., mean-based approach generates estimates that are biased and may not be consistent with the observed quantities. An alternative is the observed-value-based approach which estimates the average effect in the population. Moreover, effect-styled confidence intervals provides an alternative and a more clear way to describe uncertainty associated with the focal predictor. In addition to theoretical and methodical comparison, using simulation, we illustrate the two approaches and show that they can produce substantially different results and that observed-value-based approach can not only produce estimates consistent with the observed values, but also appropriate for bias correction. We also present an alternative way, effect-styled confidence intervals, to describe uncertainties associated with the central estimates.

4.1 Introduction

Plots of predicted values of an outcome against predictors (often called effect plots or prediction plots) are often a useful way to summarize the results of a regression model. These can be used to illustrate model uncertainty or to give a more explicit quantitative sense of how the outcome is expected to change. In generalized models with a non-linear link function, or models with a spline or polynomial response to an input variable, they can also aid in understanding difficult-to-interpret coefficient estimates (Brambor et al. 2006; Berry et al. 2012; Leeper 2017).

To make an outcome plot, we use a *focal* predictor on the x-axis and *central estimate* (predicted values) on the y-axis. The resulting plot will depend on choices we make about other (non-focal) predictor(s). In ordinary (Gaussian) linear regression, the non-focal choices may have a simple additive effect on the central estimate. However, when the focal predictor has interactions or non-linear response functions (e.g., spline or polynomial), non-focal choices can also affect the slope of the estimate. Additional challenges arise when dealing with "mixed" models, which incorporate random effects.

As noted, the outcome plots described above are often called prediction plots or effects plots. We endeavor here to make a conceptual distinction. The primary distinction between the two lies in how we describe the uncertainties around the central estimate. If our goal is to *predict* what we learn about the outcome variable by measuring the focal predictor, then we may want to capture a variety of sources of uncertainty, including that due to the intercept, focal and non-focal predictors, and random effects. Conversely, if we wish to focus on the *effect* of a focal predictor only, we might want to isolate uncertainty due to coefficients associated with that predictor. If we follow this convention, we expect effects plots to have narrower confidence intervals (CIs) than prediction plots.

The other distinction relates not to the method of calculating CIs, but to the model chosen for the plot. In general, if we want to predict based on a focal parameter, we are likely to want to fit a univariate model that only contains terms related to that predictor. If we want to know the effects of a predictor, we may want to control for covariates with a multivariate model, in order to estimate "direct" effects; leave covariates out, in order to estimate "total" effects (direct plus indirect); or take an intermediate strategy. Shi et al. (Shi et al. 2017), for example, used multivariate effects plots to visualize estimated direct effects of different predictors in a paper that compared the difference in sexual risk behaviors between circumcised and uncircumcised men.

Producing this sort of outcome plot has some challenges, including:

1. choosing the *reference point* for non-focal predictors in multivariate models
2. uncertainty estimation – appropriate choice of *anchor* for computing confidence intervals in effects plots
3. biases induced by non-linear transformations of the response variable in generalized linear models (especially generalized mixed models).

Representative values for a focal predictor are generally chosen using quantiles, or equally spaced values, for a continuous predictor; or an exhaustive set of levels for a categorical predictor. The central estimates are then calculated by holding the non-focal predictors at a reference point (a value chosen for a non-focal predictor) while varying the focal predictor, with the goal that the estimates represent how the model responds to the changes in the focal predictor (Fox and Hong 2009; Hanmer and Ozan Kalkan 2013). These values have been called: *predictor effects* (Fox and Hong 2009), *marginal predictions* (Leeper et al. 2017) or *estimated marginal means* (Lenth 2022). In this article, we refer to these quantities as the *central estimates* of the outcome.

In a model with non-focal predictors such as multivariate models, reference points can be chosen as the average of the non-focal *linear predictor variables* – we call this approach *mean-based* reference point and is currently not implemented in commonly used R software packages. We introduce an alternative choice for the reference point.

For a linear model, the averaging is done on the linear scale, i.e., linear averaging. As a result, the model-center estimates (made using the mean-based approach) are unbiased. However, in a model with non-linear link function, this is not usually true. When averaging is done on a separate link scale, the mean of the estimates is not the same as the estimate at the mean point. This leads to bias: in this case a systematic difference between the values seen on average for a given value of the focal predictor and the value predicted by the mean-based approach. An alternative to the mean-based reference point is the *observed-value-based* approach, discussed later, which involves computing the prediction over the population of non-focal predictors and then averaging across the values of the focal predictor (Hanmer and Ozan Kalkan 2013).

This article will discuss and implement various approaches for computing predictions and effects plots. We further explore and demonstrate, using simulated data, approaches for correcting bias in central estimates for generalized models involving non-linear link functions, including models with random effects. The proposed method and R software package will complement the existing ones by providing: 1) a straightforward way to generate effects plots (in our sense), and 2) a robust way to correct for non-linear averaging bias in generalized (mixed) models.

4.2 Definitions

In order to discuss the statistical background and mathematical formulation of the proposed approaches, we need to understand and formally define a number of terms, some of which have been introduced in the previous section:

Input variables Refers to the observed (or scientific) variables underlying an inference or exploration. For example, the regression models described by Eq. 4.16 and Eq. 4.17 both have 3 input variables – x_1, x_2, x_3.

Focal predictors We call the input variable on the x-axis of an outcome plot the focal predictor. Any other input variables are "non-focal" predictors.

Model matrix Refers to the design matrix whose rows include all combination of input variables. Consider an example for three hypothetical households – the first household head is a Christian with an income of $50, the second household head is Muslim with an income of $100 while the third household head is a Jew with an income of $77. Suppose we want to model the household size (`hhsize`) as function of these household characteristics, i.e.,

$$\text{hhsize} = \beta_0 + \beta_1 \times \text{income} + \beta_{2[r]} \times \text{religion} + \text{error}.$$

The model matrix corresponding to this model is given by

$$\begin{bmatrix} Intercept & income & religionJew & religionMuslim \\ 1 & 50 & 0 & 0 \\ 1 & 100 & 0 & 1 \\ 1 & 77 & 1 & 0 \end{bmatrix}.$$

The first column represents the constant term in our model, β_0. For continuous input variables, the representation in the model matrix is the same as the corresponding input variables (for example second column representing income). For categorical variables, however, by default, the model matrix creates additional dummy variables using the reference cell parameterization. This means that, if an input variable has L factor levels, then there will be $L - 1$ dummy columns representing all but the first level created in the model matrix. In our example, the column `religion` is missing and instead we have `religionJew` and `religionMuslim`; the missing category `religionChristian` is treated as the reference category.

Linear predictor variables Refer to the variables which are combined to make the linear predictor (corresponding to the columns in the model matrix). Each input variable may correspond to one or more linear predictor variables. Input variables with more than two categories (for instance religion in our previous example), or input variables with non-linear response functions (e.g., spline or polynomial) will correspond to more than one linear predictor variable – we call such multi-parameter variables (MPVs).

Model center Is the value of the central estimate calculated using focal and non-focal model center values. It is a calculation that is equal to the average of central estimates for a linear model (identity link function) without complex interactions. The average of the linear predictor variables corresponding to the focal and non-focal predictors are the focal and non-focal model center, respectively. Focal and non-focal model center values are also referred to as *center point*. For simple continuous input variables, averaging linear predictor variables is the same as averaging the input variables (see Fig 4.1).

Reference point The values (or sets of values) chosen for non-focal predictors, when estimating the predictions and effects. Typically the center point, but can also be chosen as a baseline value or as a mean across categories. We will also discuss using a set of quantiles or observations as a reference.

Anchor The value chosen for the focal predictor when estimating effect-style confidence intervals. The anchor choice does not affect the central estimates, nor prediction-style. Often chosen as the center point of the linear predictor variables corresponding to the focal predictor.

Prediction-style and effect-style plots A prediction-style plot is about predicting observations for a given value of the focal predictor. For this, we want to use the classic curved confidence intervals. An effect-style plot attempts to visualize the effect of a focal predictor and are characterized with narrower confidence intervals. Unlike a prediction-style plot, where the confidence intervals capture the uncertainty associated all predictors in the model, the effect-style plot focuses on uncertainty associated with the focal predictor only and depend on the anchor.

4.3 Statistical formulation

To estimate the central estimate of an outcome plot on the response scale, we need a link function g and appropriate values of the focal predictor $\mathbf{X}^f$ and reference point for the non-focal predictors $\mathbf{X}^{\{n\}}$. As previously mentioned, we can either use mean-based or observed-value-based reference point.

Using mean-based reference point, the central estimate is

$$\begin{aligned}\hat{\eta}_i^f &= \hat{\boldsymbol{\beta}}\mathbf{X}_c^f \\ \hat{y}_i^f &= g^{-1}(\hat{\eta}_i^f),\end{aligned} \tag{4.1}$$

where $\mathbf{X}_c^f = \{\mathbf{X}^f, \mathbf{X}_c^{\{n\}}\}$ is a centered model matrix with appropriately chosen values of the focal predictor $\mathbf{X}^f$ and centered non-focal linear predictor variables $\mathbf{X}_c^{\{n\}}$ constructed by replacing the corresponding values in the model matrix with their averages.

On the other hand, using observed-value-based reference point, Eq. 4.1 becomes

$$\begin{aligned}\hat{\eta}_j^f &= \hat{\boldsymbol{\beta}}\mathbf{X}_j^f \\ \hat{y}_i^f &= \underset{j}{\text{mean}}\ g^{-1}(\hat{\eta}_j^f), \end{aligned} \tag{4.2}$$

where $\mathbf{X}_j^f = \{X_j^f, \mathbf{X}^{\{n\}}\}$ is the model matrix of the jth observation and $\mathbf{X}^{\{n\}}$ is the entire population of the non-focal linear predictor variables. Here, we generate a vector of size $J \times N$ and then average over the values of the focal predictor; J and N represents the number of focal values and population of non-focal predictors, respectively.

For identity link function, e.g., in a simple linear model, Eq. 4.1 and Eq. 4.2 are equivalent. In the subsequent sections, we will first discuss general formulation of mean-based approach and how to generate associated confidence intervals. Thereafter, we will discuss the second approach, observed-value-based, and its application to bias correction.

4.3.1 Mean-based approach

An alternative formulation of Eq. 4.1 involves expressing the linear predictor as the sum of the focal and non-focal predictors' linear predictors. In particular,

$$\hat{\eta}_i^f(x^f, \bar{x}^{\{n\}}) = \hat{\beta}^f x^f + \sum \hat{\beta}^{\{n\}} \bar{x}^{\{n\}} \tag{4.3}$$

$$\hat{y}_i^f = g^{-1}\left(\eta_i(x^f, \bar{x}^{\{n\}})\right) \tag{4.4}$$

where x^f and $\bar{x}^{\{n\}}$ are columns of $\mathbf{X}_c^f$ corresponding to focal predictor and non-focal predictors, respectively.

In a model with MPVs or input variables with complex interactions, construction of $\hat{\eta}_i^f(x^f, \bar{x}^{\{n\}})$ is not usually straightforward since we want $\bar{x}^{\{n\}}$ to be or represent the "true" model center. We therefore illustrate some of these cases.

Dealing with multi-parameter variables

Multi-parameter variables (MPVs) such as splines, polynomials, etc., can be within the focal predictor, or within non-focal predictor(s). To distinguish the two, suppose the model which describes the hypothetical simulation of household size based on a number of socio-demographic factors such as age and wealth index is

$$\begin{aligned}\text{hh size}_i = \beta_0 + \beta_{\text{A}_1}\text{Age}_i + \beta_{\text{A}_2}\text{Age}_i^2 + \beta_{\text{A}_3}\text{Age}_i^3 \\ + \beta_{\text{W}}\text{Wealthindex}_i + \epsilon_i.\end{aligned} \tag{4.5}$$

In the first case, with `Age` as the focal predictor, is a cubic polynomial with three linear predictor variables ($\text{Age}_i, \text{Age}_i^2$ and Age_i^3). In this case, each linear predictor variable is evaluated separately across the chosen levels of the focal predictor, Age_i. Specifically, the linear predictor variables are treated as additional columns of the model matrix evaluated with the same values chosen for the focal predictor, while non-focal predictors are fixed at their reference point as discussed in the previous section. The central estimates associated with `Age` on the linear predictor scale become

$$\begin{aligned}\hat{\eta}_i^f(\text{Age}_i, \overline{\text{Wealthindex}}^{\{n\}}) = \hat{\beta}_0 + \hat{\beta}_{\text{A}_1}\text{Age}_i + \hat{\beta}_{\text{A}_2}\text{Age}_i^2 + \hat{\beta}_{\text{A}_3}\text{Age}_i^3 \\ + \hat{\beta}_{\text{W}}\overline{\text{Wealthindex}}^{\{n\}}.\end{aligned} \tag{4.6}$$

In the second case, with `Wealthindex` as the focal predictor, the non-focal predictor, `Age`, is a cubic polynomial. In this case, the non-focal linear predictor variables ($\text{Age}_i, \text{Age}_i^2$ and Age_i^3) are all treated as separate non-focal linear predictor variables and an appropriate choice of reference point applies just like in the models without MPVs. For instance, in mean-based approach, we average all non-focal MPVs. Thus

$$\begin{aligned}\hat{\eta}_i^f(\text{Wealthindex}_i, \{\overline{\text{Age}}, \overline{\text{Age}^2}, \overline{\text{Age}^2}\}^{\{n\}}) = \hat{\beta}_0 + \hat{\beta}_{\text{A}_1}\overline{\text{Age}} + \hat{\beta}_{\text{A}_2}\overline{\text{Age}^2} + \hat{\beta}_{\text{A}_3}\overline{\text{Age}^3} \\ + \hat{\beta}_{\text{W}}\text{Wealthindex}_i.\end{aligned} \tag{4.7}$$

Dealing with interactions in input variables

Interactions can be between non-focal predictors or between focal and non-focal predictors. Handling former case is similar to that of the second case in MPVs (Eq. 4.7). In the latter case, consider model described by Eq. 4.17. The interaction is between the focal, x_2, and non-focal, x_3, predictors. In this case, the values of the focal predictors are chosen as previously described and the reference point for the interacting non-focal predictor is predetermined or appropriately chosen set of values. In our example, suppose we pick i and j unique values of the focal predictor, x_2, and interacting non-focal predictor, x_3, respectively. The central estimate on linear predictor becomes

$$\hat{\eta}^f_i(x_{2i}, x_{3j}, \bar{x}_1^{\{n\}}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1^{\{n\}} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3j} + \hat{\beta}_{23} x_{2i} x_{3j}.$$

The main point is that, in the case of non-interacting non-focal predictors, the reference point is a center point while in the case of interacting non-focal predictors, the choice of the reference point is not necessarily a center point but can be any appropriate value or set of values.

4.3.2 Uncertainty estimation

We describe the uncertainty around the estimates using confidence intervals (CIs). In principle, every prediction has a different CI. The conventional way to compute variances for predictions is

$$\sigma_i^2 = \mathrm{Diag}(\mathbf{X}^{\star}\mathbf{\Sigma}\mathbf{X}^{\star\top}), \tag{4.8}$$

so that the confidence intervals are $\hat{\eta}^f_i \pm q\sigma_i$, where $\mathbf{\Sigma}$ is the variance matrix of $\hat{\boldsymbol{\beta}}$ and q is an appropriate quantile of Normal or t distribution (Fox and Hong 2009). This generates conventional CIs which incorporate all the uncertainties – including the uncertainties due to the intercept and non-focal predictor. We call this prediction-style CIs.

Effect-style CIs

But what if we are interested in the uncertainty as a result of the focal predictor only (effect-style CIs), so that the CIs are $\hat{\eta}_i^f \pm q\sigma_i^f$, i.e., effects? For effect-style CIs, we need an anchor and centered model matrix with all non-focal linear predictor variables set to zero.

Let $\mathbf{X}_c^f$ be a centered model matrix previously defined, and let $\mathbf{A}$ be an anchor matrix, with the same dimensions and entries of all non-focal linear predictor variables as $\mathbf{X}_c^f$. Let $\mathbf{A}^f$ be the column of $\mathbf{A}$ corresponding to focal linear predictor variable defined in $\mathbf{X}_c^f$. Any appropriate values can be chosen for $\mathbf{A}^f$ but for model center (center-anchored), we use $\mathbf{A}^{\bar{f}}$ which is the mean of the focal predictor. Thus

$$\boldsymbol{\sigma}_i^2 = \mathrm{Diag}((\mathbf{X}_c^f - \mathbf{A}^{\bar{f}})\boldsymbol{\Sigma}(\mathbf{X}_c^f - \mathbf{A}^{\bar{f}})^\top). \tag{4.9}$$

We can see that $\forall\ \mathbf{X}^f = \mathbf{A}^f$, $\mathbf{X}_c^f - \mathbf{A}^f = \mathbf{0}$, hence $\boldsymbol{\sigma}_i^2 = \mathbf{0}$. Similarly, for all values of $\mathbf{X}^f$ close to anchor point $\mathbf{A}^{\bar{f}}$, the term $(\mathbf{X}_c^f - \mathbf{A}^{\bar{f}})$ and $\boldsymbol{\sigma}_i^2$ goes to $\mathbf{0}$ and $\boldsymbol{\sigma}_i^2 = \mathbf{0}$ if $\mathbf{X}^f = \mathbf{A}^{\bar{f}}$. This means that $\boldsymbol{\sigma}_i^2$ close to the anchor point are smaller than those away from the anchor point; and results to confidence intervals which are narrower around the anchor or crosses at the anchor point for simple models. In other words, this shows the effect of changing the focal value from the anchor value. By setting $\mathbf{A}^f = \mathbf{0}$, we get the variances for the prediction-style CIs in Eq. 4.8.

An alternative way to compute $\boldsymbol{\sigma}_i^2$ in Eq. 4.9 is by *zeroing-out* the covariance matrix, which involves setting all the non-focal linear predictor variables in Σ to 0. This procedure can be implemented in commonly used R packages for effect plots and prediction plots, but only works when the input variables are *centered* prior to model fitting, in case of numerical variables, and more complicated when the input variables are categorical. The anchor approach, however, does not require the input variables to be centered prior to model fitting since the computation of $\mathbf{X}_c^f - \mathbf{A}^{\bar{f}}$ affects only the intercepts and non-focal predictor linear variables – the slopes and variance corresponding to the focal predictor linear variables are not affected.

4.4 Bias correction

When dealing with non-linear link functions and additional non-focal predictors, generated estimates may not reflect the observed response due to the bias, i.e., a term we use to describe a situation in which the central curve does not align well with the observed data points – the central curve is either below or above majority of the observed data points. In such cases, bias correction is needed when back-transforming the estimates to the original scales. The common approach for bias-adjustment is second-order Taylor approximation (Duursma and Robinson 2003; Hanmer and Ozan Kalkan 2013); already implemented in **emmeans** (Lenth 2022). Here, we describe and implement a different approach – observed-value-based approach for bias correction. Hanmer and Kalkan (Hanmer and Ozan Kalkan 2013) discussed and implemented observed-value-based and input-variable-based mean-based approaches in binary response models only. Our formulation is more general and can easily be extended to other link functions other than logistic.

We can directly compare the outcome estimate at the model center for non-focal parameters (mean-based estimate) with the average of predictions evaluated across the population values of non-focal predictors (observed-value-based estimate, the colon below indicates that we are comparing two quantities):

$$g^{-1}\left(\eta_i^f(\bar{x}^f, \bar{x}^{\{n\}})\right) : \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(\eta_i^f(\bar{x}^f, x^{\{n\}})\right). \tag{4.10}$$

In an ordinary linear model, the link function g is the identity function, so the two means are the same. For non-trivial link functions, we expect them to be different in general.

From Jensen's inequality, the exponential link function, for example, is concave up; hence we expect the right-hand side of Eq. 4.10 to be greater than the left-hand side, i.e., the average of the predictions is greater than the prediction at the average of the focal and non-focal parameters. On the other hand, the logistic function is concave up and concave down at low and high probabilities, respectively. We expect a pattern similar to that of exponential function when the logistic function is concave up and the opposite when the logistic function is concave down. In other words, the mean-based approach can under-estimate the prediction in exponential and low probability logistic functions and over-estimate in high probability logistic function.

4.4.1 Observed-value-based approach for bias correction

An alternative approach to choosing a reference point is to compute central estimates over all observations of the non-focal predictors (members of the population) (Hanmer and Ozan Kalkan 2013). The non-linear transformation involved in these computations is always *one-dimensional*; all of the multivariate computations required are at the stage of collapsing the multidimensional set of predictors for some subset of the population to a one-dimensional distribution of $\hat{\eta}_i^f(x_f, x_{\{n\}})$, which is a function of the chosen values of the focal predictor and the whole sample of non-focal predictors, as opposed to the definition in Eq. 4.3. More specifically:

- compute linear predictor associated with whole sample of the non-focal predictors, $\hat{\eta}_j^{\{n\}} = \sum \hat{\beta}^{\{n\}} x^{\{n\}}$
- compute linear predictor associated with the chosen values focal predictor, $\hat{\eta}_i^f = \hat{\beta}^f x^f$
- for every value of the focal linear predictor, $\hat{\eta}_i^f$, compute

$$\begin{aligned}\hat{\eta}_j^f(\hat{\eta}_i^f, \hat{\eta}_j^{\{n\}}) &= \hat{\eta}_i^f + \hat{\eta}_j^{\{n\}} \\ &= \hat{\eta}_j^f(x^f, x^{\{n\}}).\end{aligned} \tag{4.11}$$

Once Eq. 4.11 is computed, we back-transform the estimates to the original scale and average over the levels of the focal predictors, j:

$$\hat{y}_i^f = \underset{j}{\text{mean}}\ g^{-1}\left(\hat{\eta}_j^f(x^f, x^{\{n\}})\right). \tag{4.12}$$

We make similar adjustments to compute the variances of the predictions at every level of the focal predictor:

$$\sigma_j^2 = \mathrm{Diag}(\mathbf{X}_j^f \boldsymbol{\Sigma} \mathbf{X}_j^{f^\top}) \tag{4.13}$$

and

$$\mathrm{CI}_i = \operatorname*{mean}_j \; g^{-1}\left(\hat{\eta}_j^f(x^f, x^{\{n\}}) \pm q\sigma_j^f\right). \tag{4.14}$$

We make further adjustments for models with random effects components to correct the bias induced by the random effects components. We treat the random effects components as additional non-focal predictors in the observed-value approach and modify Eq. 4.11. In particular

$$\begin{aligned} \tau^{\{n\}} &= \mathbf{Z}b \\ \hat{\eta}_j^f(x^f, x^{\{n\}}, \tau^{\{n\}}) &= \hat{\eta}_j^f(x^f, x^{\{n\}}) + \tau^{\{n\}} \end{aligned} \tag{4.15}$$

where $\mathbf{Z}$ and b are the design matrix and a vector of random effects, respectively.

4.5 Mean-based vs. observed-value-based

In the observed-value-based approach, the ensemble of predictions and CIs are back-transformed before averaging, see Eq. 4.12, so we do not need to worry about the non-linear averaging. In other words, the averaging is no longer on the link scale and is likely to be bounded by the original data scale. In simple linear models without interactions, averaging on the link scale is identical to averaging on the response, so both approaches yield similar results. However, picking a single value, e.g., the mean of the predictor, on which to draw conclusions about the effect can be problematic, unrealistic, or not contained in or representative of the population. In addition, the mean-based approach fails to use every value of non-focal predictors hence not utilizing the full potential of the information contained in the data. This may limit the inferences we can make about the entire population. In general, the mean-based approach provides the predictions of an average case, whereas the observed-value-based approach summarizes the predictions over the entire population. In some applications, the effect of an average case might not be generalizable to the entire population, especially if the average does not represent the population. This might not be a problem in the observed-value-based approach since it focuses on specific observations – the prediction is first obtained for each observation and then averaged across the levels of the focal predictor.

Another potential concern with the mean-based approach arises when direct naive use leads to a rare or meaningless basis for generalization. For example, if our sample has 20% Jews, 30% Muslims, and 50% Christians. One approach (default in common packages) is assigning equal category weight, i.e., 1/3 Jews, 1/3 Muslims, and 1/3 Christians (the "sum-to-zero" approach). The second approach (our default) is setting dummy categorical variables in the model matrix to their means, which, by default, set them to their sample means or observed proportions. The first or second approach translates to prediction for a household head who is 1/3 or 50% Christian, respectively (Hanmer and Ozan Kalkan 2013). The second approach seems more realistic and will converge to the population mean in many cases.

The observed-value-based approach is not entirely foolproof. For instance, similar to the mean-based approach, in the case of continuous focal predictors, choosing the representative values of the focal predictors can be very challenging, especially if the cases are not evenly distributed around the minimum and the maximum values or within some subgroups defined in the population. In addition, the observed-value-based approach can be computationally intensive for large datasets.

4.6 Simulation examples

We start by comparing our proposed approach with the existing implementations and then illustrate the construction of outcome plots. We also demonstrate that the mean-based approach works well for interactions and MPVs, and that the observed-value approach can correct the bias induced by non-linear averaging.

4.6.1 Comparison with other implementations

The most commonly used R software packages for outcome plots (**emmeans** and **effects**), by default, use the average of input variables and "sum-to-zero" as the reference point approach for continuous and categorical input variables, respectively. However, there are a number of choices one can make when constructing outcome plots – for example, in the presence of interactions, the default for **emmeans** and **effects** is to average the input variables and use these averaged values for the interaction, as opposed to our preferred model-center approach (averaging each linear predictor variable separately). The packages give the same results as our method for models without interactions. However, when interactions are present, the two approaches can produce substantially different results, with the model-center approach more closely matching the observed values.

To illustrate this, we simulate data from the models below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{4.16}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{23} x_2 x_3 + \epsilon. \tag{4.17}$$

We simulate using $\beta_0 = 5$, $\beta_1 = -3$, $\beta_2 = 1$, $\beta_3 = 2$, $\beta_{23} = 5$, $\{x_{1,2,3}, \epsilon\} \sim \text{Normal}(0, 1)$, and then compare estimates from the **emmeans**, **effects** and our proposed alternative (**varpred**) to the "true" central estimates (calculated using the simulation parameters) and mean of the data, i.e., $\bar{y}$, as shown in Fig 4.1.

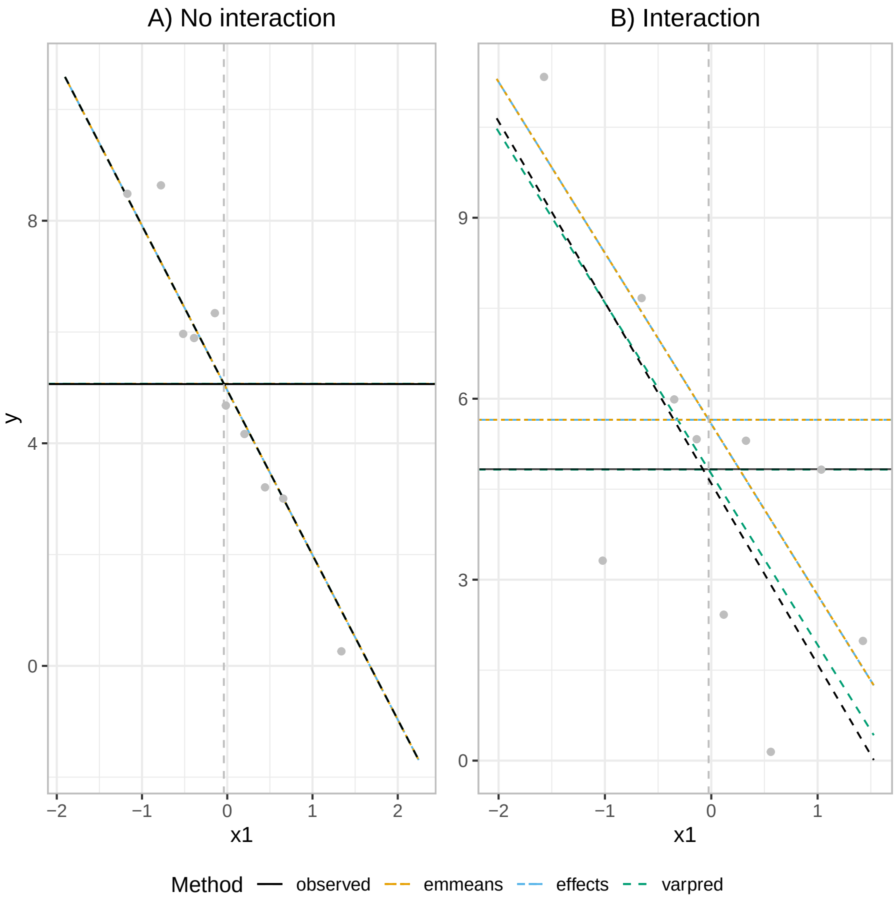


Figure 4.1: **A comparison of central estimates for emmeans, effects and varpred.** The horizontal dashed lines are the mean of the central estimates and simulated y, $\bar{\hat{y}}$ and $\bar{y}$, respectively. The grey points are averages of the simulated points binned according to the value of x_1. The trend lines represent the $\hat{y}$ – central estimate. A: In the absence of interaction, the predicted mean, $\bar{\hat{y}}$, is the same in all three approaches and closely matches the actual mean (truth); the central estimate likewise matches the truth (horizontal lines). B: With a simple interaction between non-focal predictors, results from **emmeans** and **effects**, but not from the proposed **varpred**, are biased

In the absence of interactions (Eq. 4.16), the three approaches produce identical estimates, which match the simulated values, Fig 4.1A. However, the estimates start to differ in the presence of interactions, even as simple as the one in Eq. 4.17. In particular, estimates from **emmeans** and **effects** are identical but differ from **varpred**'s, which is very close to the simulated average ($\bar{y}$), Fig 4.1B. In the simple model, Fig 4.1A, the input variables are the same as the linear predictor variables, so all the three methods produce identical results. In the interaction model, Fig 4.1B, there is an additional linear predictor variable $(x_2 x_3)$. **emmeans** and **effects** first average the input variables to compute $\bar{x_2}\bar{x_3}$ while **varpred** first calculates the corresponding vector of linear predictor values and then averages.

Our implementation, **varpred**, can generate both prediction-style and effect-style plots. However, as previously mentioned, it is hard to generate an effect-style plot in **emmeans** and **effects**. We consider Eq. 4.16 and use x_2 as the focal predictor, and compare prediction and effect plots using **varpred**. We also compare different anchors for the effect plot. See Fig 4.2.

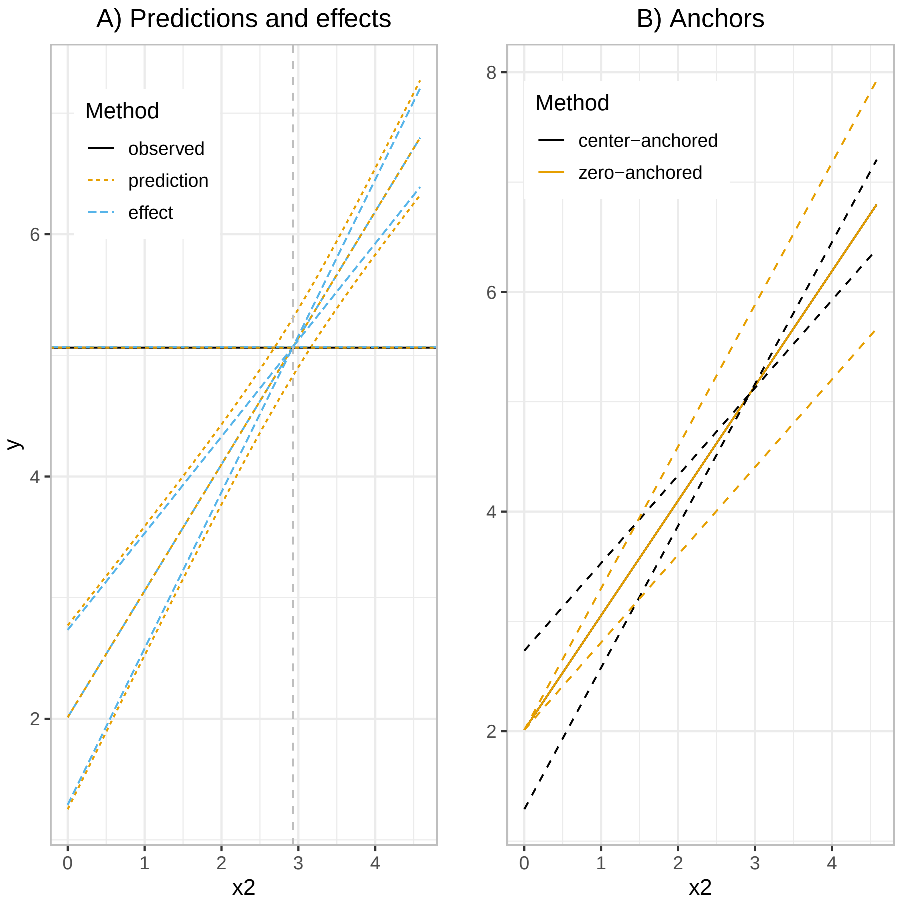


Figure 4.2: **Prediction and effect plots.** The description of horizontal, vertical, and trend lines remain the same as above. A: The wider dashed curves correspond to the conventional prediction curves, while the narrower curves crossing at the center point represent the effect from **varpred**. For simple OLS models, the effect-style curves cross at the center point (center-anchored, the default). The prediction-style curves incorporate the uncertainties due to the intercept term and other non-focal predictor, while effect-style curves only consider the focal predictor's main effect uncertainty. B: center- and zero- anchored effects. The zero-anchored effect means that the anchor is at the zero value of the focal predictor. The choice of the anchor does not affect the central estimates.

For a prediction-style plot, the confidence intervals are wider because they 301
include uncertainties associated with the intercept and non-focal predictors but 302
are narrower and cross at the mean of the focal predictor. In other words, with 303
varpred, we can generate effects indicating uncertainty due only to *changes* in the 304

focal predictor (thus in the case of a simple focal predictor, there is no uncertainty at the anchor point, in this case, the model center).

4.6.2 Prediction- and effect-style plots for MPVs

To illustrate the distinction between predictions and effects in MPVs, we simulated a multivariate model described in S1 Appendix. We also used the same model to illustrate how predictions and effects differ if MPVs are in the focal or non-focal predictor. In particular, we generated two sets of mean-based predictions and effects: 1) one with `age` as the focal predictor (a cubic polynomial - MPV); and 2) one with `Wealthindex` as the focal predictor (linear). We used the center point as the anchor for the effects in both cases, as shown in Fig 4.3.

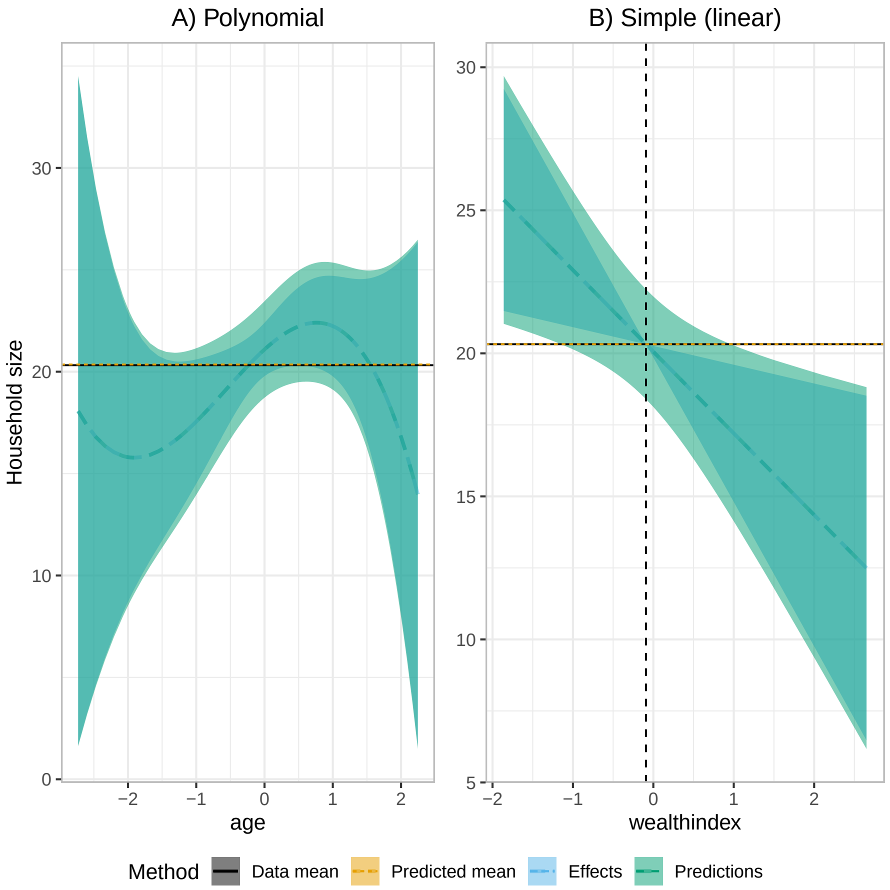


FIGURE 4.3: **Central estimate, prediction and effect plots for cubic polynomial and simple focal predictors.** A: The focal predictor is a MPV cubic polynomial. B: The focal predictor is linear, with MPV non-focal predictor. The central estimates (dashed central curves or lines) are the same for predictions and effects in both cases. We use the model center as the anchor for effect-style CIs. For a simple predictor (panel B), this corresponds simply to the mean of the input variable; thus, the effect curves intersect at that point. For an MPV predictor (panel A), the center point does not correspond to a single value of the predictor, and the curves do not cross. The horizontal black and yellow dashed lines are observed and predicted average household size, i.e., $\overline{hh\ size}$ and $\overline{\widehat{hh\ size}}$, respectively.

We expect the CIs for effect-style plots to cross at the anchor point in a model with a simple focal predictor. On the other hand, if the focal predictor is an MPV, the center point is not expected to correspond to a single value of the focal predictor. In this case, the effects will be narrower than the predictions plots but will not necessarily intersect. For linear models, we expect mean of the data, the average of the central estimates and the prediction at the model center to all be identical.

4.6.3 Bias correction

We simulated data motivated by the water, sanitation, and hygiene (WaSH) study in which we were interested in investing the contribution of demographic and socio-economic factors to improved WaSH indicators among slum dwellers in Nairobi, Kenya. In this particular study, we used the mean-based approach to generate the predicted probabilities. However, we noticed that the predictions consistently over- or under- estimated the observed proportions; and did not align well with the observed data points. To demonstrate this, we consider a binary-outcome simulation with two input variables (described in S2 Appendix), such that `Age` has a very small effect size in comparison to `Wealthindex`, and compared their effects on the estimated probability of improved water quality as shown in Fig 4.4.

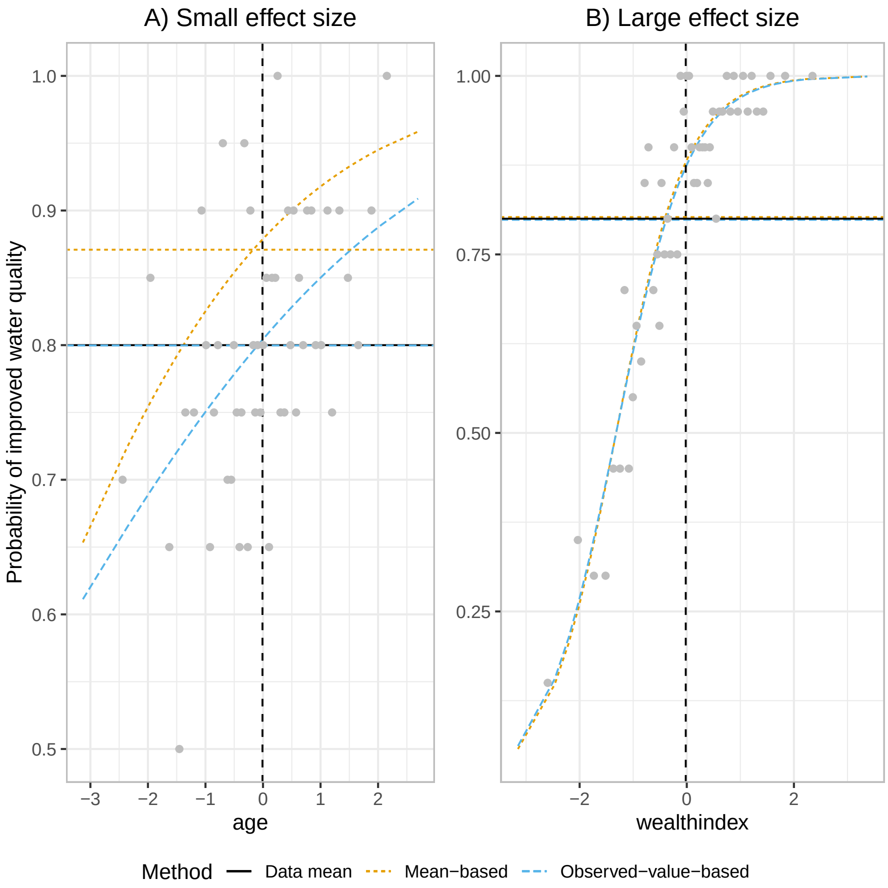


Figure 4.4: **Mean-based and observed-value-based central estimates.** If the focal predictor has a small effect size, mean-based and observed-value-based approaches produce different estimates – A. However, in the case of strong effect size, the two approaches produce very close estimates – B. The difference in mean-based and observed-value-based is due to the bias induced by the non-focal predictor and the non-linear averaging in the logistic model. The mean-based approach is affected by the non-linear averaging. If the focal predictor has a small effect size, the bias is even pronounced since the effect of non-linear averaging is primarily driven by the non-focal predictor(s) with a strong effect. Since the observed-value-based approach averages over the whole population of the non-focal linear predictor variables, it accounts for the effects of non-linear averaging. The horizontal dashed lines correspond to the respective averages. The vertical dashed lines represent the center point, and the point at which they intersect the horizontal lines represents the expected “perfect” estimate at the model center. The grey points are binned observations – observed proportions of improved water quality in each bin.

If there were no effect of non-linear averaging, then we would expect the average observed proportion and the average predictions to intersect at the center point as we see in Fig 4.3B. One possible reason for the variations we see in Fig 4.4 is the non-linear averaging; since both observed status and predicted probabilities are averaged on the response scale as opposed to the link scale. For example, if the range of values is bigger than 0.5 (seemingly the case here), then we would expect the averages to be slightly higher than we would expect at the center point.

In Fig 4.5, we use the results in Fig 4.4A and compare the central estimate, prediction and effect plots associated with the central estimates.

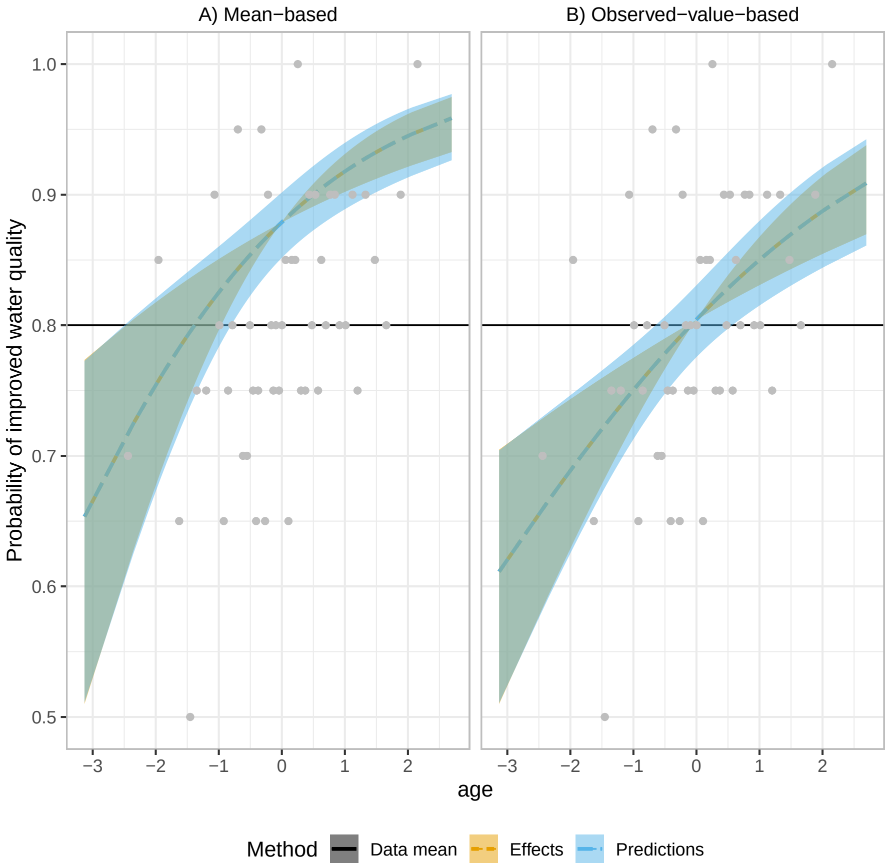
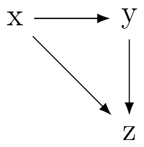


FIGURE 4.5: **Bias correction for central, effect and prediction estimates.** As in Fig 4.4, the observed-value based estimates match the pattern of the data better than mean-based estimates.

4.6.4 Mediated effect

Mediated effects provide a good example of some of the choices involved in showing an effect-style plot. To illustrate this, consider an example in which age of the head of the household (`x`) has a direct on the availability of improved water service in the household (`z`), as well as an indirect effect through its effect on the income of the head of the household (`y`), as illustrated below:



The arrows above show the direction of influence, and the implication of $x \to y$ and $y \to z$ is that x affects z in two ways. First, it has direct effect; second, it can have an indirect effect by influencing y, regulated by parameter ρ as shown in S3 Appendix.

To make inferences about this mediated effect, we can ask questions about the *total* and *direct* effects. In the first case, we want to fit a univariate model which regresses z on x, excluding y. We refer to this as *non-mediated* model. In the second case, we want to fit a multivariate model which regresses z on x and y. We refer to this as *mediated* model. We simulated the data as shown in S3 Appendix and fitted the two models – non-mediated and mediated. We compared mean-based and the observed-value-based central estimates in both cases, as shown in Fig 4.6.

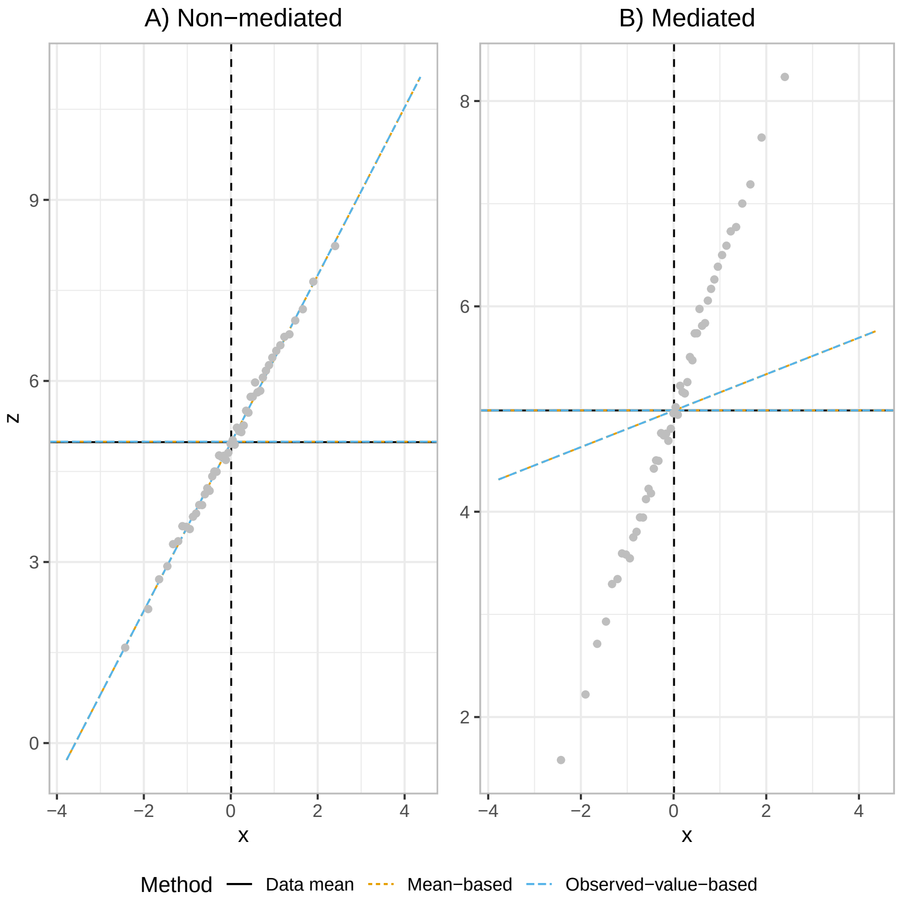


Figure 4.6: **Mean-based and observed-value-based estimates for non-mediated (total) and mediated (direct) effect models.** A: In the absence of a mediator variable, the central estimate aligns well with the observed data. B: In the presence of indirect effect (mediator variable included), the central estimates do not align with the observations. Since the models in A and B are linear models with identity link functions, both approaches (mean-based and observed-value) give the same estimates. The horizontal dashed lines are the respective average estimates and mean of the data ($\bar{\hat{z}} \approx \bar{z}$). The vertical dashed black represents the center point, and the point at which it crosses the horizontal lines represents the expected "perfect" estimate at the model center. The grey points are the binned observations.

From Fig 4.6A, we see what we would expect if we simulated a univariate model with no non-focal predictors, even though in the simulation, the effect of x on z is mediated through y. This is the total effect of x, which only tells us the influence of x on z. By ignoring y in the model, we can still capture the effect of x and match the observed values using both approaches. However, if we control for the mediator variable y, the estimates do not necessarily match the observed values since x indirectly affects z. By controlling for the mediator variable, we can get an indication of the direct (very weak) effect of x on z (see Fig 4.6B).

4.7 Discussion and conclusion

Generalized linear (mixed) models are widely used in various fields, including public health. In a model involving difficult-to-interpret coefficient estimates, an outcome plot can aid in understanding and summarizing the results. In particular, a prediction plot would be appropriate if the goal is to capture every uncertainty in the model for a particular focal predictor or if we are interested in total effect. Conversely, an effect plot is preferable if we want to focus on the uncertainty associated with a focal predictor only or if we are interested in the direct effect.

The mean-based approach is widely used to create outcome plots. However, in a model with complex interaction, MPVs or categorical variables, it is sensitive to the choice of the reference point. We have demonstrated that a model-center-based reference point is generally a stable choice and provides estimates more consistent with the observed quantities as compared to common input-variable mean-based approach.

In a model with a non-linear link function such as a logistic or exponential function, the generated central estimate curve may not match well with the observed data, i.e., our description for bias. In such a model, the observed-value-based approach provides a way to generate more consistent estimates and is preferable to the widely used mean-based approach.

The argument and results we present in this paper support a greater need for a shift in focus on how to summarize these kinds of models. From our theoretical, methodological and simulation results, researchers using these models should, in the absence of theoretical justification, report predictions based on the observed-value approach or at least attempt to compare the two approaches before settling on the most appropriate in answering their research question. Moreover, we provide R package, **varpred**, which implements these methods and is available on GitHub (https://cygubicko.github.io/varpred/).

Our simulation examples focused on simple linear and logistic models due to their wide range of usage and application. These models also act as a starting point for building other complex models, including mixed effect models and models with categorical predictors. The logic for extending to more complex models, including other forms of non-linear link functions, is straightforward. The components needed for extension are the correct linear predictor and the inverse link function; everything else generalizes. In addition, our R package implementation already extends to and supports most of the non-linear link functions and mixed model framework, including multivariate binary outcome models.

4.8 Supporting information

S1 Appendix. Cubic polynomial interaction simulation. Consider a hypothetical simulation which simulates household size as a function of household wealth index and cubic function of the age of the household head, specified as follows:

$$\begin{aligned}
\text{hh size}_i &= \beta_0 + \beta_{\mathrm{A}_1}\text{Age}_i + \beta_{\mathrm{A}_2}\text{Age}_i^2 + \beta_{\mathrm{A}_3}\text{Age}_i^3 + \beta_{\mathrm{W}}\text{Wealthindex}_i + \epsilon_i \\
\text{Age}_i &\sim \text{Normal}(0,1) \\
\text{Wealthindex}_i &\sim \text{Normal}(0,1) \\
\epsilon_i &\sim \text{Normal}(0,10) \\
\beta_0 &= 20 \\
\beta_{\mathrm{A}_1} &= 0.1 \\
\beta_{\mathrm{A}_2} &= 0.8 \\
\beta_{\mathrm{A}_3} &= 0.3 \\
\beta_{\mathrm{W}} &= -0.5 \\
i &= 1, \cdots, 100 \qquad (4.18)
\end{aligned}$$

S2 Appendix. Binary outcome simulation. Consider a simple simulation for improved water quality in Nairobi slums, such that the status is 1 for improved and 0 for unimproved water quality. In addition to the focal predictor, age of the household head, we add wealth index. In particular:

$$\begin{aligned}
\text{status}_i &\sim \text{Bern}(\text{P}_\text{i}) \\
\text{logit}(\text{P}_\text{i}) &= \eta_i \\
\eta_i &= \beta_0 + \beta_\text{A}\text{Age}_i + \beta_\text{W}\text{Wealthindex}_i \\
\text{Age}_i &\sim \text{Normal}(0, 1) \\
\text{Wealthindex}_i &\sim \text{Normal}(0, 1) \\
\beta_0 &= 5 \\
\beta_\text{A} &= 0.5 \\
\beta_\text{W} &= 1.5 \\
i &= 1, \cdots, 10000
\end{aligned} \tag{4.19}$$

S3 Appendix. Mediated effect simulation. Next, we consider a simple indirect mediation previously described and simulate a binary outcome model such that:

$$\begin{aligned}
z_i &= \beta_0 + \beta_{xz}x_i + \beta_{yz}y_i \\
y_i &= \rho x_i + \sqrt{1 - \rho^2} y_y \\
x_i &\sim \text{Normal}(0, 1) \\
y_y &\sim \text{Normal}(0, 1) \\
\rho &= 0.8 \\
\beta_0 &= 5 \\
\beta_{xz} &= 0.2 \\
\beta_{yz} &= 1.5 \\
i &= 1, \cdots, 10000
\end{aligned} \tag{4.20}$$

Acknowledgments

This work was supported by a grant to Jonathan Dushoff from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery.

Author Contributions

Conceptualization: Jonathan Dushoff, Steve Cygu

Software: Steve Cygu, Benjamin M. Bolker

Writing – original draft: Steve Cygu

Chapter 5

A multivariate longitudinal analysis with binary outcomes: Correlates of sanitary improvements in the slums of Nairobi

The text I present here is a draft of a manuscript planned for submission for publication

Author Contributions

SC performed statistical analyses with helpful feedback from JD , DTK, and BMB; SC wrote the first draft of the manuscript, and all authors revised the manuscript.

A multivariate longitudinal analysis with binary outcomes: Correlates of sanitary improvements in the slums of Nairobi

Steve Cygu[1], Benjamin M. Bolker[1,2], Damazo T. Kadengye[3], Jonathan Dushoff[1,2]

[1]School of Computational Science and Engineering, McMaster University, Hamilton, Ontario, Canada
[2]Department of Biology, McMaster University, Hamilton, Ontario, Canada
[3]African Population and Health Research Center, APHRC Campus, Manga Close, Off Kirawa Road, P.O. Box 10787-00100, Nairobi, Kenya

* cygubicko@gmail.com

Abstract

Access to improved water, sanitation, and hygiene (WaSH) services remain a challenge to many households living in informal urban settlements in Kenya. To understand the temporal and household level dynamics between the WaSH services, using data from Nairobi Urban and Demographic Surveillance System (NUHDSS), this study employs separate (univariate) and joint (multivariate) outcome models for binary outcomes to investigate the contribution of demographic and economic factors to WaSH access in two informal urban settlements in Nairobi, Kenya, namely Korogocho and Viwandani. The results showed a few differences, but no any generalizable patterns, between the univariate and multivariate outcome models in terms of estimated effect sizes, prediction, and effect plots. However, the multivariate outcome model estimated additional quantities – household and year level correlation between the services. For example, at household level, the status of toilet and garbage disposal services tend to move in the opposite direction, i.e., households with improved water services are likelier to have improved toilet and unimproved garbage disposal services. Most importantly, the result points to the need for researchers and policymakers to consider the possibility of correlation between WaSH indicators in ways beyond those explained by the explanatory variables.

5.1 Introduction

Many residents of urban areas face multiple obstacles while accessing basic services. These obstacles can be socio-economic, institutional, spatial, or political barriers; and are even more prevalent and severe in informal urban settlements (Pierce 2017). In Kenya, the Nairobi Urban Health and Demographic Surveillance System (NUHDSS) has collected data since 2003 on household-level access to water, sanitation, and hygiene (WaSH) services in two major slums in Nairobi, Kenya (Beguy et al. 2015). The NUHDSS data provides a useful starting point for understanding the factors associated with trends in access to WaSH services in slum areas.

Like other developing African countries, Kenya has experienced unprecedented urban growth and urbanization, which has adversely affected the quality of life and left many urban populations with a huge unmet demand for basic services. Nearly two-thirds of urban residents have no access to improved sanitation (Chikozho et al. 2019).

Even though WaSH services constitute some of the most basic requirements for human health and dignity, they are either inadequate or unavailable in most Nairobi slums (Chikozho et al. 2019). Several interventions have been intended to upgrade most slums in Nairobi, focusing on infrastructural development issues, especially in the Korogocho and Viwandani slums. However, these communities still face various challenges, including overcrowding, poverty, alcohol, and crime-related issues. In addition to widespread poverty, residents of these slums face near-absence of most of the basic services they need to live healthy lives (Chikozho et al. 2019; Iddi et al. 2021).

Our main study outcome is measured by three WaSH variables (drinking water, toilet facilities, and garbage disposal). This is a challenging dataset for a number of reasons. Specifically, we expect outcome variables to be correlated, beyond factors explained by our predictors, due to causal interactions and unmeasured covariates. We also expect temporal correlations between measurements from each household. Dealing with these considerations is complicated because the outcome variables are binary, and we thus cannot use standard techniques based on Gaussian responses.

We want a modeling approach that can: 1) control for household correlations, 2) address correlation between outcomes, and 3) estimate outcome-specific effects. One way to address 1 and 2 would be by combining outcome indicators into scores (representing an aggregated assessment of outcomes for each household) and then regressing the scores against the covariates. This method fails to address 3; because it only shows how predictors connect to the aggregate outcome, not with particular outcomes. Univariate-response models – with separate mixed logistic linear models being fitted for each outcome, for example – may address 3 at the expense of 1 and 2. Specifically, separate models will ignore the fact that the outcomes observed from the same subject (household) are likely to be correlated since they are subject to shared influences that are distinctive to that particular household (LaLonde et al. 2019). Ignoring such correlations may lead to poor estimates (Miro-Quesada et al. 2004; Ivanova et al. 2016; Fang et al. 2018; LaLonde et al. 2019). Alternatively, we can attempt to address all three goals by jointly modeling the three outcomes (Fang et al. 2018; LaLonde et al. 2019).

There are several potential advantages to this joint-modeling approach. First, due to its joint formulation, both outcome-specific and global effects can be estimated (LaLonde et al. 2019). Second, the association between outcomes can be captured in terms of correlation between household-level random effects (Ivanova et al. 2016). Third, joint modeling may increase statistical power. However, as noted above, correctly formulating and fitting these models is particularly challenging in our case of binary outcomes.

Prior studies have used NUHDSS data to explore WaSH indicators in informal urban settlements. Chikozho et al. 2019 used a generalized estimating equation (GEE) to model each of the three WaSH indicators. Iddi et al. 2021 applied multistate transition models to assess the effect of socio-economic factors on each of the WaSH indicators. Although these two models accounted for the household-level correlation between years, they did not address potential correlation between WaSH services.

This study aimed to apply and validate a joint modeling approach to analyze binary outcomes. Our approach took into account the longitudinal nature of the data to model all three WaSH outcome variables (improved water, toilet facilities, and garbage disposal). The analysis was based on a generalized linear mixed model approach. We compared separate (univariate) and joint (multivariate) outcome models based on simulated and NUHDSS WaSH data. By investigating the contribution of demographic and economic factors to WaSH access, we hope that the identified, sometimes overlooked, factors will be helpful to the government and other development workers seeking to extend and equalize access.

5.2 Methodology

5.2.1 Data acquisition and study design

We used data from a longitudinal NUHDSS covering two major urban slums, Korogocho and Viwandani, in Nairobi, Kenya. The baseline survey that defined the initial population for the NUHDSS was carried out between July–August 2002. Subsequently, data from demographic, socio-economic, household characteristics, and livelihood sources were collected yearly until 2015. We used a subset of data from 2006 to 2015 because most of the covariates considered for the analysis were collected from 2006. Beguy et al. 2015 give a complete description of the NUHDSS study design and setting; Iddi et al. 2021 provide a summary of the household characteristics of the study population.

WaSH outcomes

We considered three WaSH variables: Drinking water source, toilet facility type, and garbage disposal method. Each of these is classified as improved or unimproved – this follows WHO guidelines adapted by Chikozho et al. 2019 and Iddi et al. 2021 from Yu et al. 2016. Table 5.1 provides a summary of the classification scheme.

Table 5.1: Classification of WaSH indicators (adapted by Chikozho et al. 2019 and Iddi et al. 2021 from Yu et al. 2016)

	Improved	**Unimproved**
Drinking water source	• Piped water into dwelling, plot or yard • Public tap or standpipe • Tube well or borehole • Protected dug well with hand pump • Protected spring • Rainwater collection from the roof	• Unprotected dug well • Unprotected spring • Small water vendor (cart with small tank or drum) • Bottled water • Tanker truck • Rainwater collection from surface run off • Protected dug well with bucket
Toilet facility type	• Flush, pour flush to piped sewer system, septic tank or pit latrine • VIP latrine • Pit latrine with slab • Composting toilet	• Flush or pour flush to elsewhere e.g., to open drain • Pit latrine without slab (slab with holes) or open pit • Bucket • Hanging toilet or hanging latrine • No facilities, bush or field
Garbage disposal method	• Garbage dump • Private pits • Public pits • Proper garbage disposal services • Other organized groups such as the national youth service	• In the river • On the road, railway line or station • In drainage, sewage or trench • Vacant, abandoned house, plot or field • No designated place or all over • Street boys or urchins • Burning • Other

Table 5.2 shows the descriptive statistics of the three WaSH outcomes. After data cleaning, there were 53491 unique households, of which many were interviewed more than once, for a total of 151730 observations, of which 90.1%, 19.3% and 47.5% showed improved water, toilet and garbage disposal services, respectively.

Table 5.2: WaSH outcome variables

	Overall
	(N=151730)
Water source	
Unimproved	15066 (9.9%)
Improved	136664 (90.1%)
Toilet type	
Unimproved	122438 (80.7%)
Improved	29292 (19.3%)
Garbage disposal	
Unimproved	79602 (52.5%)
Improved	72128 (47.5%)

There were 53491 unique households

Demographic and socio-economic explanatory variables

Table 5.3 shows the distribution and summary statistics of the demographic and socio-economic variables used as explanatory variables in the analysis. The number of interviews increased from 2006 to peak in 2011 and then dropped to the lowest value in 2015. The majority (61.7%) of the respondents were from Viwandani. The median age of the head of the household was 33 years; 77.9% of households were reported to be male-headed; median household size was 3 members.

Table 5.3: Demographic and socio-economic explanatory variables

	Overall
	(N=151730)
Interview year	
2006	8820 (5.8%)
2007	14727 (9.7%)
2008	15650 (10.3%)
2009	17917 (11.8%)
2010	19303 (12.7%)
2011	20283 (13.4%)
2012	18092 (11.9%)
2013	18550 (12.2%)
2014	17696 (11.7%)
2015	692 (0.5%)

Table 5.3: Demographic and socio-economic explanatory variables *(continued)*

	Overall
Slum	
korogocho	58074 (38.3%)
viwandani	93656 (61.7%)
Age	
Mean (SD)	35.29 (10.76)
Median [Min, Max]	33.00 [18.00, 70.00]
Gender	
female	33605 (22.1%)
male	118125 (77.9%)
Household size	
Mean (SD)	3.16 (2.04)
Median [Min, Max]	3.00 [1.00, 9.00]
Income (KSh.)	
<1,000	616 (0.4%)
1,000-2,499	1959 (1.3%)
2,500-4,999	16106 (10.6%)
5,000-7,499	36209 (23.9%)
7,500-9,999	36180 (23.8%)
10,000-14,999	35506 (23.4%)
15,000-19,999	16514 (10.9%)
20,000+	8640 (5.7%)
Expenditure (KSh.)	
Mean (SD)	3338.86 (1682.44)
Median [Min, Max]	2970.00 [80.00, 9990.00]
House ownership	
Not owned	140722 (92.7%)
Owned	11008 (7.3%)
Amount of food	
Didn't have enough	53112 (35.0%)
Had enough	98618 (65.0%)
Experienced household shocks	
No	139184 (91.7%)
Yes	12546 (8.3%)
Self-rating	
Mean (SD)	3.94 (1.34)

Table 5.3: Demographic and socio-economic explanatory variables *(continued)*

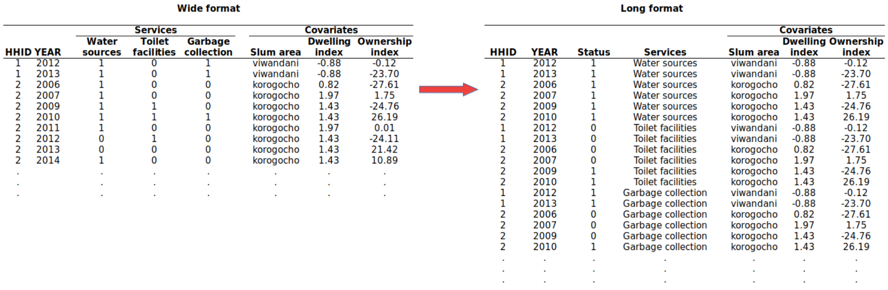
	Overall
Median [Min, Max]	4.00 [1.00, 10.00]
Dwelling index	
Mean (SD)	0.00 (1.19)
Median [Min, Max]	0.04 [-4.36, 2.19]
Ownership index (within)	
Mean (SD)	0.00 (1.72)
Median [Min, Max]	-0.25 [-8.68, 10.77]
Ownership index (outside)	
Mean (SD)	0.00 (2.33)
Median [Min, Max]	0.19 [-2.68, 12.41]

5.2.2 Data pre-processing

Joint outcome variable

Currently, the well-known frequentist frameworks (such as packages **lme4** and **glmmTMB**) in R software used for this kind of data do not have a natural syntax for multivariate outcomes. Consequently, we restructured the data to long format by generating a new categorical variable, `Services`, whose categories are the old service types (water, toilet, and garbage disposal services), and recoded the outcome variables to a single `Status` variable as shown in Table 5.4. The new structure treats `Services` as a "repeated measure" from a particular household in a particular year. It is used as an interaction for the model's fixed and random terms.

Table 5.4: The structure of multivariate data in long format.



Wide format

		Services			Covariates		
HHID	YEAR	Water sources	Toilet facilities	Garbage collection	Slum area	Dwelling index	Ownership index
1	2012	1	0	1	viwandani	-0.88	-0.12
1	2013	1	0	1	viwandani	-0.88	-23.70
2	2006	1	0	0	korogocho	0.82	-27.61
2	2007	1	0	0	korogocho	1.97	1.75
2	2009	1	1	0	korogocho	1.43	-24.76
2	2010	1	1	1	korogocho	1.43	26.19
2	2011	1	0	0	korogocho	1.97	0.01
2	2012	0	1	0	korogocho	1.43	-24.11
2	2013	0	0	0	korogocho	1.43	21.42
2	2014	1	0	0	korogocho	1.43	10.89
.		.	.	.	.	.	.
.		.	.	.	.	.	.
.		.	.	.	.	.	.

Long format

				Covariates		
HHID	YEAR	Status	Services	Slum area	Dwelling index	Ownership index
1	2012	1	Water sources	viwandani	-0.88	-0.12
1	2013	1	Water sources	viwandani	-0.88	-23.70
2	2006	1	Water sources	korogocho	0.82	-27.61
2	2007	1	Water sources	korogocho	1.97	1.75
2	2009	1	Water sources	korogocho	1.43	-24.76
2	2010	1	Water sources	korogocho	1.43	26.19
1	2012	0	Toilet facilities	viwandani	-0.88	-0.12
1	2013	0	Toilet facilities	viwandani	-0.88	-23.70
2	2006	0	Toilet facilities	korogocho	0.82	-27.61
2	2007	0	Toilet facilities	korogocho	1.97	1.75
2	2009	1	Toilet facilities	korogocho	1.43	-24.76
2	2010	1	Toilet facilities	korogocho	1.43	26.19
1	2012	1	Garbage collection	viwandani	-0.88	-0.12
1	2013	1	Garbage collection	viwandani	-0.88	-23.70
2	2006	0	Garbage collection	korogocho	0.82	-27.61
2	2007	0	Garbage collection	korogocho	1.97	1.75
2	2009	0	Garbage collection	korogocho	1.43	-24.76
2	2010	1	Garbage collection	korogocho	1.43	26.19
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Predictors

Several data pre-processing steps were undertaken to prepare the dataset for modeling. In particular; `age` was scaled (mean centered and divided by the standard deviation); `household size` was log transformed; `interview year` was scaled; `self-rating`, households' rank on a scale of 1 (poorest) - 10 (richest), was scaled; `household shocks`, a binary variable which indicated whether the household had experienced fire, eviction, rape, floods, demolition, stabbing, mugging, severe illness, lay-off, theft or death, was created; `income index` was created by ranking the income categories, the ranks were then scaled; `dwelling index` was generated from principal component analysis (PCA) of scores for various indicators of floor, lighting and wall materials; `ownership index (within or outside)` which indicated whether households owned various household assets (where they resided or elsewhere) at the time of interview was generated from PCA; `expenditure` which referred to the total amount spent by households, in the past 7 days, on food, energy, water and rent was scaled; and lastly, we created the `service status in the previous year` variable which had the following categories; *base year* for the first observation year, *improved* if the being predicted on each line service was improved in the previous observational year, *not observed* if status in the previous year was missing and *unimproved* if the service was unimproved in the previous year.

5.2.3 Statistical analysis

We are interested in modeling multivariate (correlated) binary outcomes (three response variables). One strategy, described by Chib 1998, is to use a multivariate probit model based on modeling the underlying latent variables assumed to arise from a multivariate normal distribution. While this is a convenient way of modeling multivariate binary outcomes, its parameters are difficult to interpret. Instead, for this analysis, we adopt the multivariate logistic regression model, which allows marginal distributions to follow a logistic distribution (O'Brien and Dunson 2004); the parameters of this model can be easily translated to odds ratio.

Let $\mathrm{Y}_{\mathrm{hy,s}}$ be the observed value of service (taking a value of 1 if improved) s in household h and year y. Then, each observed outcome is distributed as $\mathrm{Y}_{\mathrm{hy,s}} \sim$

$\mathrm{Bern}(\pi_{\mathrm{hy,s}})$, where $\pi_{\mathrm{hy,s}}$ is the probability that household h has improved service s in year y. Let $\mathbf{X}$ denote a matrix of covariates (demographic and socio-economic factors, plus the lagged service indicator) to be included in the model. Random variation is incorporated into the model via

$$\begin{aligned}
\tau_{\mathrm{s}}^{(0)} &= \beta_{\mathrm{s}}^{(0)} + \delta^{(\mathrm{y})} + \delta^{(\mathrm{h})} \\
\delta^{(\mathrm{y})} &\sim \mathrm{MVN}\left(0, \begin{bmatrix} \sigma_{\mathrm{W}}^2 & & \\ \sigma_{\mathrm{WT}} & \sigma_{\mathrm{T}}^2 & \\ \sigma_{\mathrm{WG}} & \sigma_{\mathrm{TG}} & \sigma_{\mathrm{G}}^2 \end{bmatrix}\right) \\
\delta^{(\mathrm{h})} &\sim \mathrm{MVN}\left(0, \begin{bmatrix} \sigma_{\mathrm{W}}^2 & & \\ \sigma_{\mathrm{WT}} & \sigma_{\mathrm{T}}^2 & \\ \sigma_{\mathrm{WG}} & \sigma_{\mathrm{TG}} & \sigma_{\mathrm{G}}^2 \end{bmatrix}\right).
\end{aligned}$$

Here $\delta^{(\mathrm{y})}$ and $\delta^{(\mathrm{h})}$ are the between-year and between-household variance of the y^{th} and h^{th} year and household among the services, respectively. The full joint model is defined as:

$$\mathrm{logit}(\pi_{\mathrm{s}}) = \tau_{\mathrm{s}}^{(0)} + \mathbf{X}\boldsymbol{\beta}_{\mathrm{s}} \tag{5.1}$$

where $\mathrm{s} \in \{\mathrm{water(W)}, \mathrm{toilet(T)}, \mathrm{garbage(G)}\}$ services, $\mathrm{logit}(\pi_{\mathrm{s}}) = \log(\pi_{\mathrm{s}}/(1-\pi_{\mathrm{s}}))$ is a vector in the range $(-\infty, \infty)$.

We fitted a generalized mixed model, i.e., a joint-outcome hierarchical logistic regression model with shared random effects that accounts for the household and year variations using R package **glmmTMB**. In particular, the random effects specification accounted for the variation between the three outcomes of the same household in a particular year and captured the unobserved factors specific to each household (in a particular year) that may influence the three services; thus allowing us to estimate the correlation between the services. We also fitted separate univariate-outcome models for each of the three outcomes, and compared these two sets of models based on coefficients and outcome plots.

5.2.4 Simulation study

In order to understand the two (separate and joint) modeling approaches, we performed simulation-based validation. The simulation was used to validate and refine the proposed models before applying them to the real WaSH dataset.

We based this simulation study on the WaSH data described in the previous section, with a slightly different but mathematically equivalent formulation of Equation 5.1, and simulated two outcome variables – status (improved or unimproved) of water and toilet services. More specifically, we simulated a multivariate binary outcome model with intercept-only random effects for household and year to jointly investigate the effect of covariates (age of head of household, wealth index, and income) on the predicted probabilities of improved services. Income had an indirect effect on the services mediated through the wealth index.

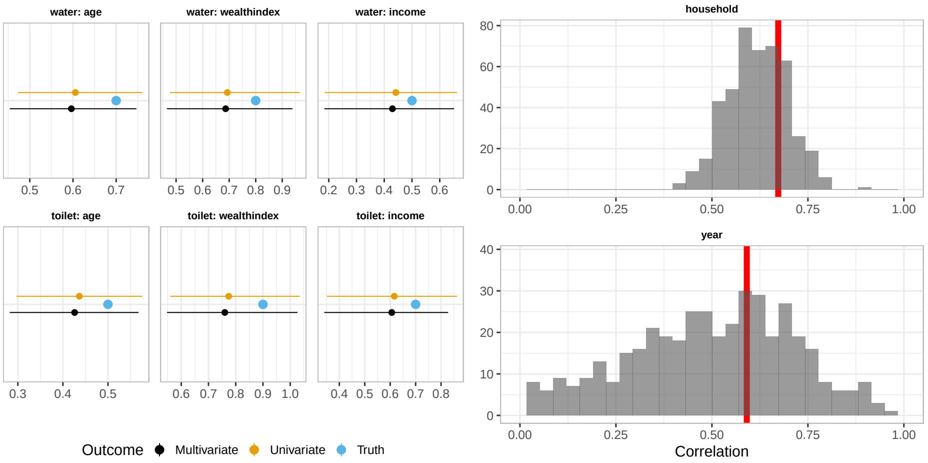
All simulations and analysis were performed in R software.

5.3 Results

In this section, we first present the results from the simulation study and then present results from the observed WaSH data.

5.3.1 Simulated data

Figure 5.1 compares the results from multivariate and univariate outcome models based on the simulated data described above. Figure 5.1a shows the 2.5%, 50%, and 97.5% quantiles of the estimated coefficients from 500 simulations and compares them to the true values used to generate the random data. Figure 5.1b shows the distribution of the correlation between the services estimated from the joint model. Generally, the estimated coefficients successfully captured the true parameters, and the joint model does not appear to indicate any bias.



(A) Coefficient estimates

(B) Correlation

Figure 5.1: *(a)* Compares the lower, median, and upper quantiles of the estimated coefficients for the joint (multivariate) and the separate (univariate) outcome models from 500 simulations to the true coefficient values. The two models generally capture the true coefficients and are not very different. *(b)* Shows the distribution of household and year level water and toilet services correlation estimates, a by-product of the multivariate outcome model. A high correlation between toilets and water at the household level means that households with improved water services are likely to have improved toilet services and vice versa. The vertical red line is the true correlation.

Figure 5.2 compares outcome plots for each predictor for one particular simulation. They show how the predicted probabilities of improved water and toilet services change at various levels of the predictor of interest, together with uncertainty associated with the predictions (effect plots based on observed-value approach for bias correction). We perform two comparisons: 1) effect plots for multivariate and univariate models, and 2) effect plots generated from the two models and the true central estimate. The two approaches substantially captured the expected patterns and did not result in any clear differences.

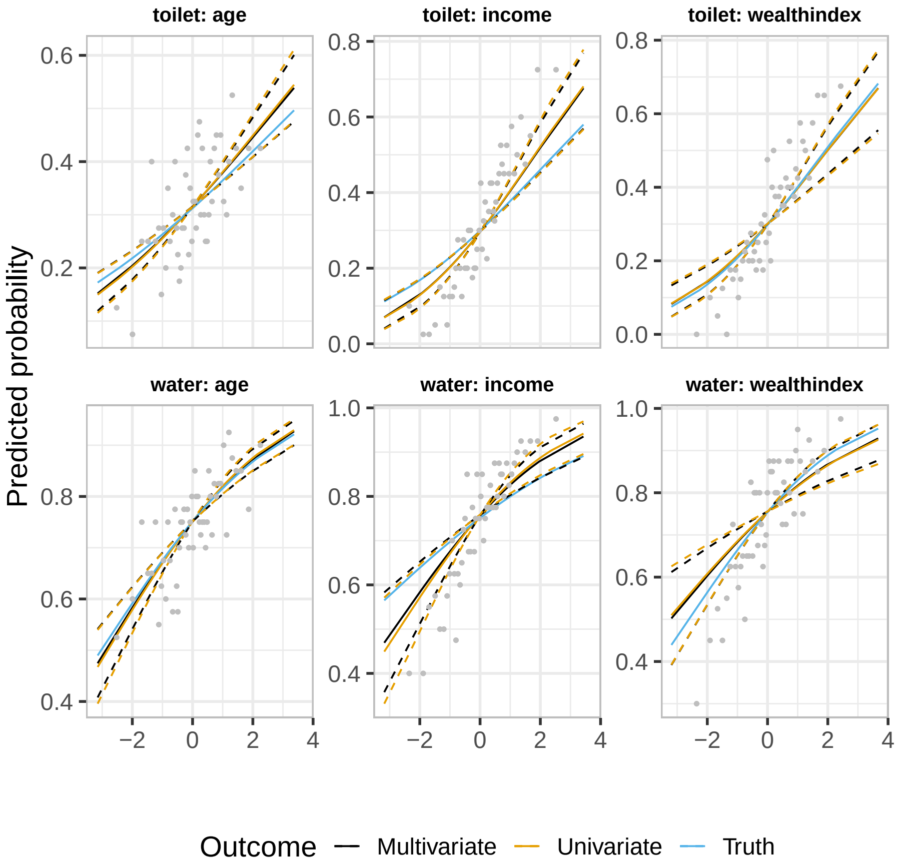
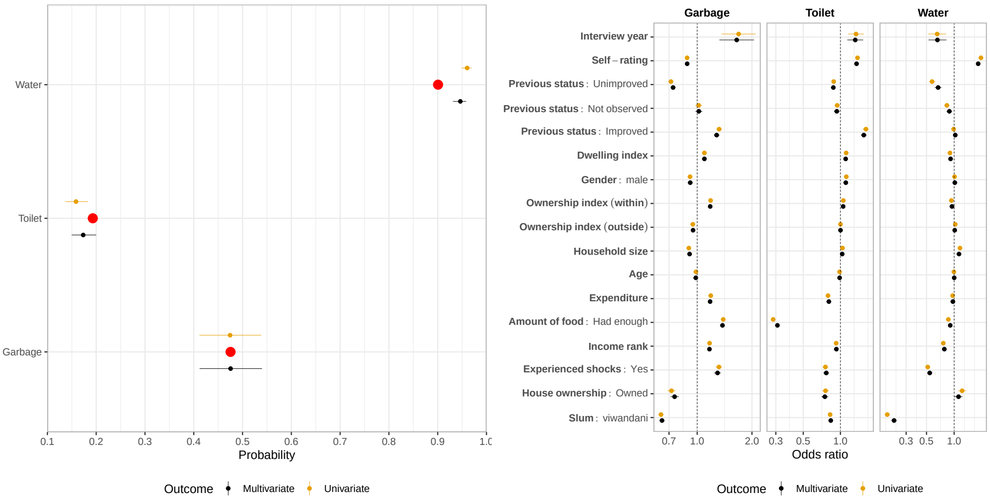


Figure 5.2: Effect plots for the multivariate and univariate outcome models for one particular simulation. The central solid curves are the central estimates; they tell us what we expect the probability of improved water or toilet services to be at a particular value of the predictor of interest. The dashed curves represent effects at the 95% interval. If the predictor has a strong effect, we expect central solid (black and orange) curves to closely match the truth (blue solid curve), have narrow effects, and align well with the simulated data (grey points). However, in the presence of a mediated effect, the estimates do not necessarily align with the observed data – as we observe in the case of income. Generally, the two modeling approaches give similar central estimate and effect patterns.

5.3.2 WaSH data

This section presents the results of the two models based on WaSH data. Our focus is to use effect plots to make inferences about the model parameters. However, we also present coefficient plots because they are commonly used for these kinds of studies. Figure 5.3 compares the baseline probabilities, i.e., the probability of improved service for an average household, taking all other factors into consideration; and odds ratios for the coefficients, together with the corresponding 95% confidence intervals, for the multivariate and univariate outcome models. Figure 5.3a compares the probability of improved services at baseline. For instance, an average household has over 90%, 50% and 30% probability of having improved water, garbage disposal, and toilet services, respectively, at the baseline. Figure 5.3b shows the odds ratios corresponding to the main effects.



(A) Baseline probabilities

(B) Main effects

Figure 5.3: *(a)* The baseline probabilities of improved services for an average household compared to the observed proportions (red points). Households are likelier to have improved water services, about equally likely to have unimproved or improved garbage disposal services, and less likely to have improved toilet services. *(b)* Compares the odds ratios corresponding to the main effects. For example, the odds of having improved garbage and toilet services increase with a unit increase in the interview year but decrease in the case of water services. Household heads who are highly rated (as rich) are associated with higher odds of an improved toilet and water services but lower odds of improved garbage disposal services. Households that had unimproved garbage disposal services in the previous year have lower odds of improving than in the base year (baseline category). A similar pattern is observed for toilet services but almost equally likely or unlikely for water services. **Bold** categories with ":" are categorical predictors with the following baseline category: previous status (base year), gender (female), amount of food (did not have enough), house ownership (rented), experienced shocks (no) and slum (Korogocho).

The estimated baseline probabilities in Figure 5.3a over-estimate the relatively high proportion of improved drinking water, and under-estimate the relatively low proportion of improved toilet facilities, as expected due to non-linear averaging across sources of variation. Figure 5.4 shows effect plots with an *observed-value-based* bias correction.

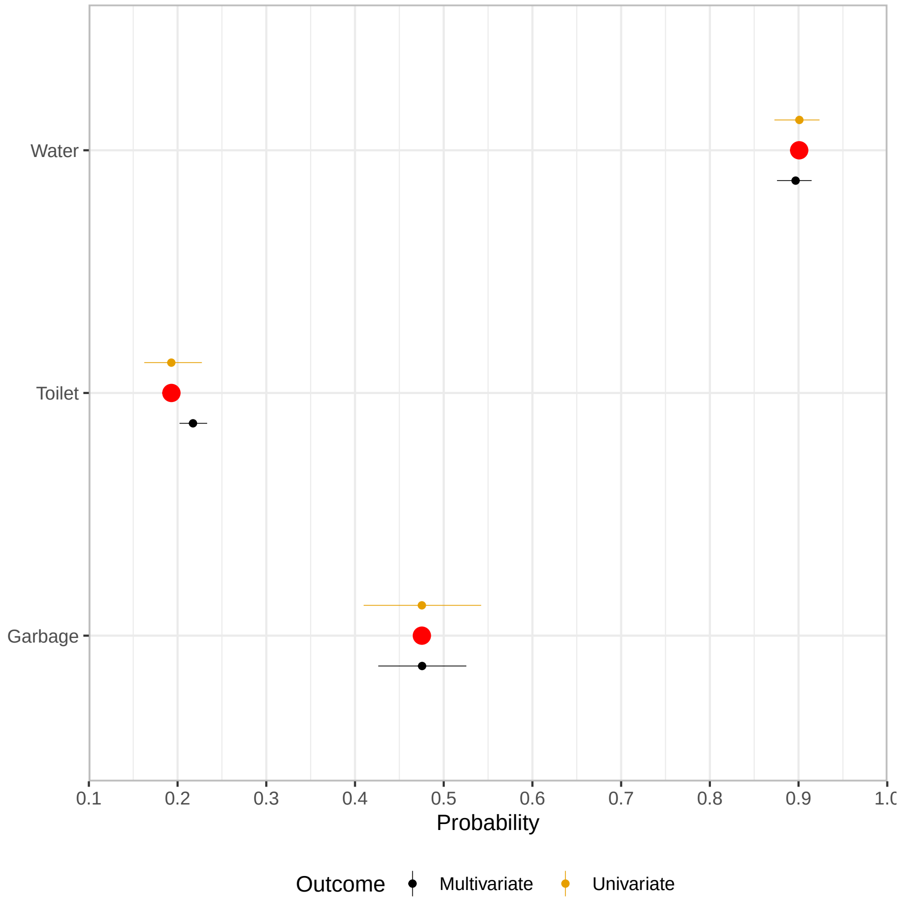


Figure 5.4: Effect plot for the baseline probabilities of improved services, corrected for bias due to non-linear averaging. The estimates closely match the observed proportions (red points).

Figure 5.5, 5.6 and 5.7 compares effect plots for the two models.

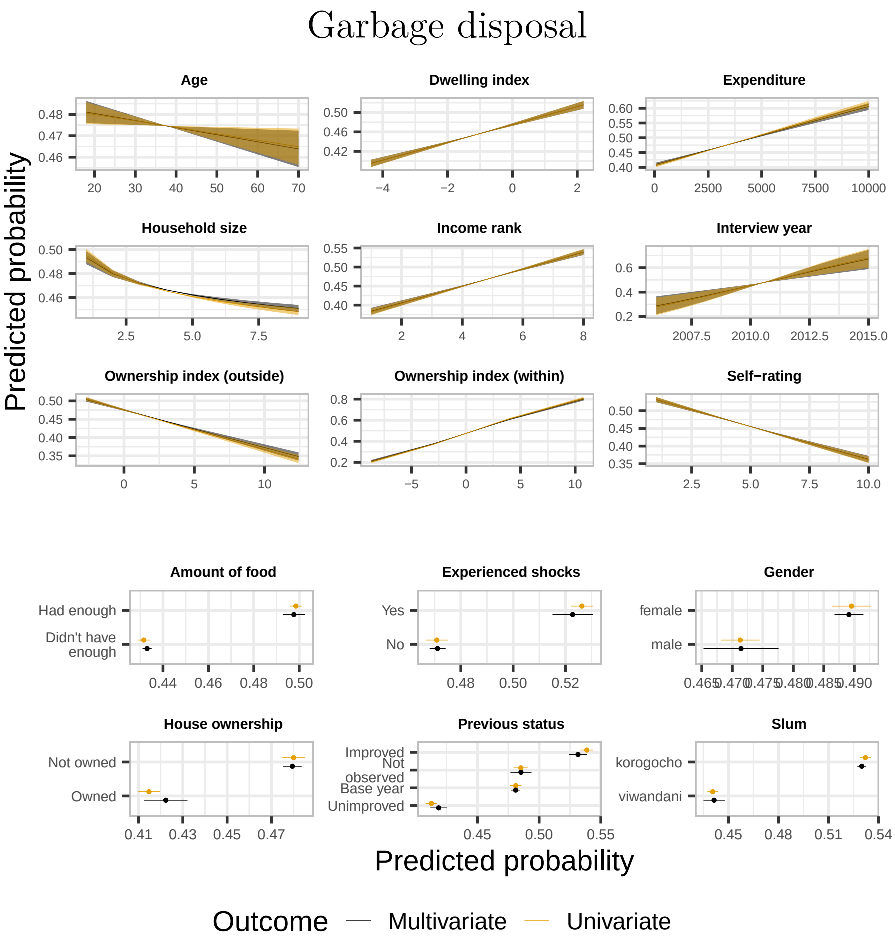


Figure 5.5: Effect plots for garbage disposal comparing multivariate and univariate outcome models. Households with older head of household have a low probability of improved garbage disposal. Similar patterns are observed in large households, households with a high index for items owned outside the households and highly rated households (rated as rich). On the other hand, households that spent more have a higher probability of improved garbage disposal services. Similarly, households with high income, which were interviewed in recent years, had a high dwelling index and a high index for items owned within the households have a higher probability of improved garbage disposal services. Further, households that had enough to eat, had experienced at least one household shock, were female-headed, did not own the house they lived in, had improved garbage services in the previous year, and were from Korogocho have a higher probability of improved garbage services than those in the corresponding category. There are no clear differences in the estimates between the two modeling approaches.

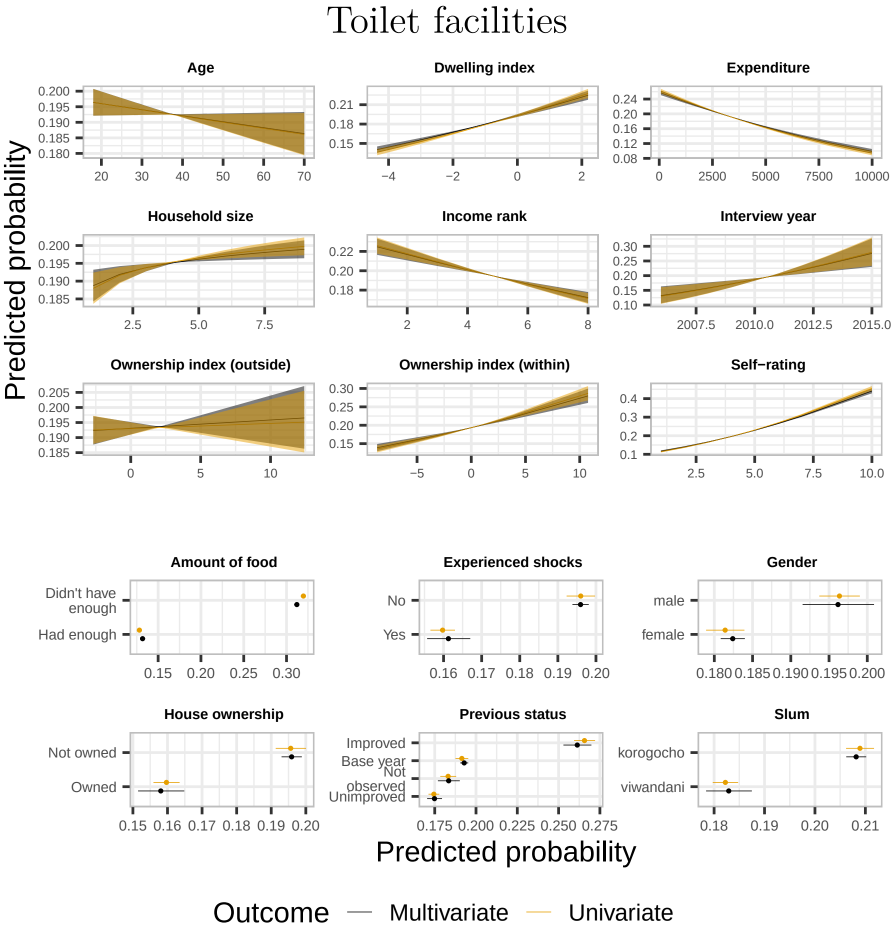


Figure 5.6: Effect plots for toilet facilities comparing multivariate and univariate outcome models. Households with older head of household that spent more and had high incomes are associated with a low probability of improved toilet services. On the other hand, large households, households which had a high index for items owned within the households, households interviewed in recent years, households that had a high dwelling index, and highly rated households are associated with a higher probability of improved toilet services. Further, households that did not have enough to eat, had not experienced at least one of the household shocks, were male-headed, did not own the house they lived in, had improved toilet services in the previous year, and were from Korogocho have a higher probability of improved toilet services. The effect of the ownership index (items owned away from the place of residence) is unclear, as indicated by the almost-flat central estimate line and wide effect plots. We observe few (in household size and ownership index) differences in the estimates between the two modeling approaches.

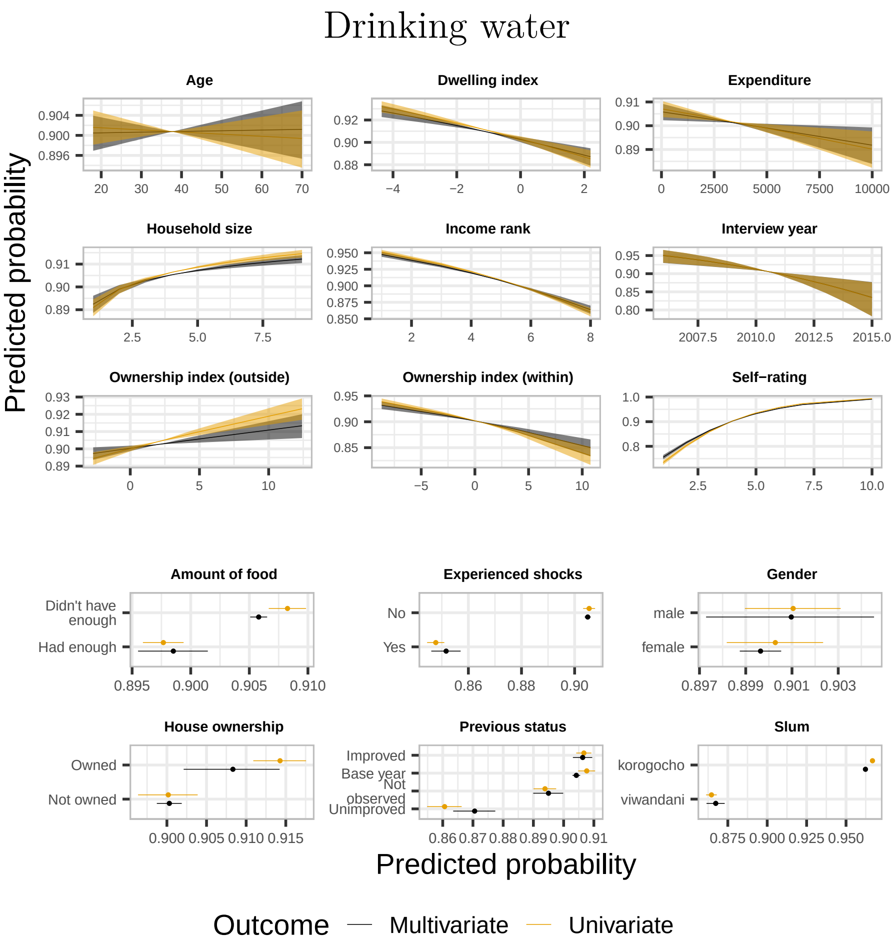


Figure 5.7: Effect plots for water services comparing multivariate and univariate outcome models. Households that spent more, had higher income, were interviewed in recent years, had a high dwelling index, and had high ownership index (items owned within the households) are associated with a low probability of improved water services. On the other hand, large households, households that had high ownership index for items owned away from where they resided and highly rated households are associated with a high probability of improved water services. Further, households that did not have enough to eat, had not experienced at least one of the household shocks, owned the house, and were from Korogocho are associated with a higher probability of improved water services. However, there are no clear differences between female and male-headed households and the status of the improved water service in the previous year. Also, the effect of the age of the head of household is unclear, as indicated by the almost-flat central estimate line and wide effect plots. We also observe a few noticeable differences between the two approaches, particularly age and ownership-related predictors.

In Figure 5.8, we narrow down and compare predictions for the three outcomes from the multivariate outcome model on the same scale.

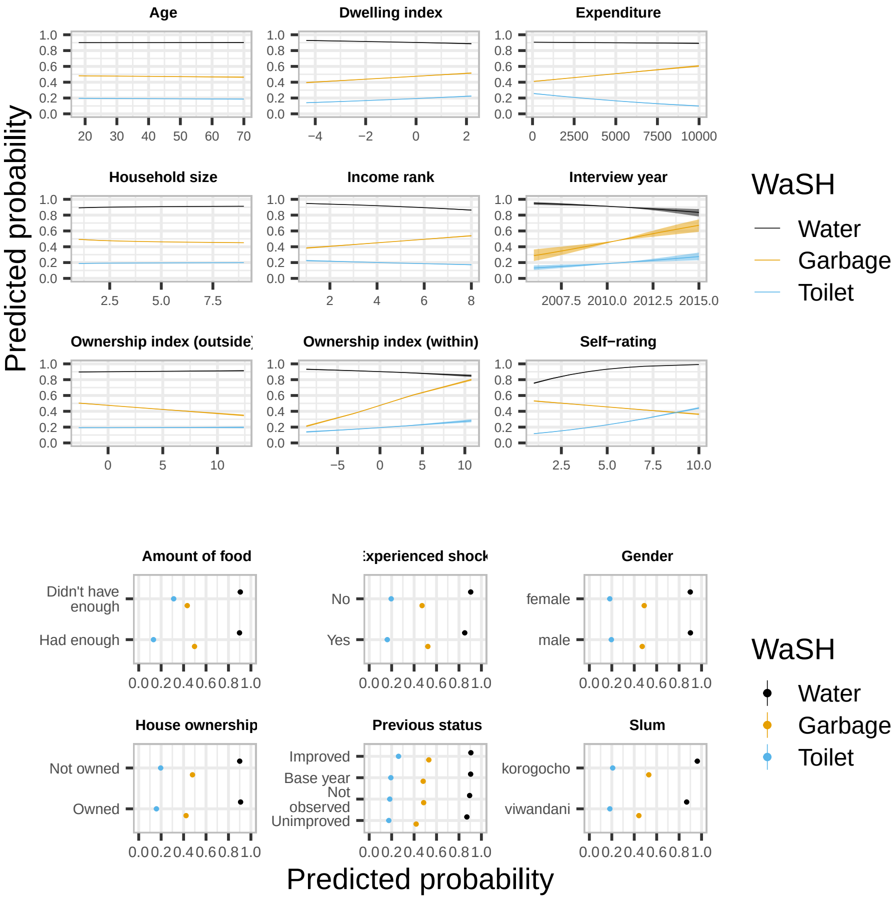


Figure 5.8: A comparison of effect plots for the multivariate outcome model. For all three outcomes, households with older head of household are associated with a lower probability of improved services. Households that had a high dwelling index, were recently interviewed and had a high ownership index (items owned within the households) are associated with a higher probability of improved garbage disposal and toilet but not water services. Households that spent and earned more are associated with a higher probability of improved garbage disposal service but not toilet and water services. Larger households, households that had a high ownership index (items owned away from the place of residence) and were highly rated (as rich), are associated with a slightly higher probability of improved water and toilet services but not garbage disposal services. Households that had enough food and were headed by a female are associated higher probability of improved garbage disposal services but not water and toilet services. Households that owned the residential house are associated with a higher probability of improved water services but not garbage disposal and toilet services. Households from Korogocho, and households that had improved services in the previous year are associated with a higher probability of improved service across all three services.

The joint modeling approach (multiple outcome) provides a way to estimate the household and year-level correlation between the outcomes, as shown in Figure 5.9.

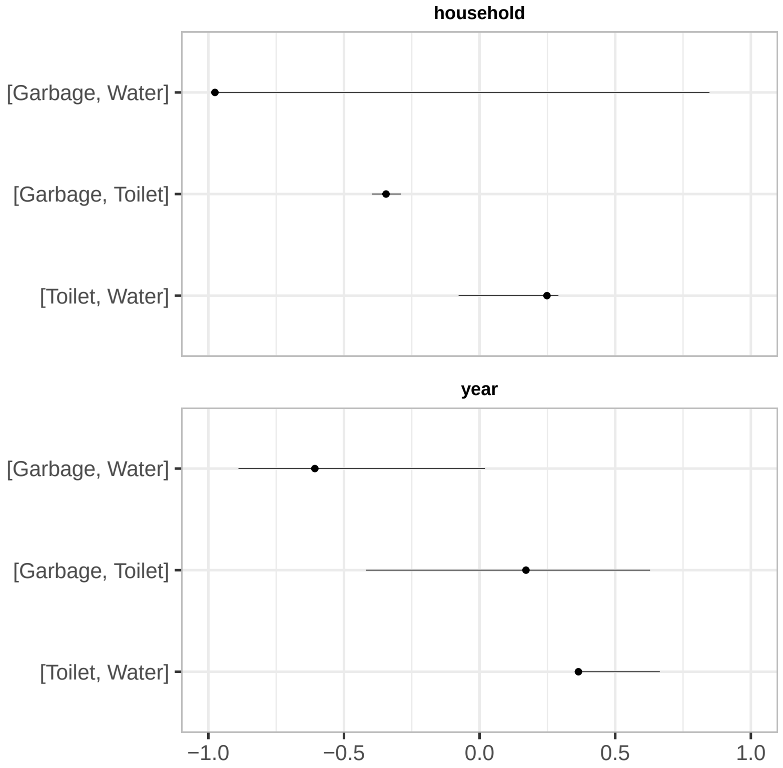


FIGURE 5.9: Correlation (together with the Wald confidence intervals) between the services at household and year level. Households with improved garbage disposal services are less likely to have improved toilet or water services. On the other hand, households with improved toilet services are more likely to have improved water services. Garbage disposal and water services are likely to have opposite patterns within the households and throughout the years. Toilet and water services are likely to have similar trends across the years.

5.4 Discussion

In Kenya, informal urban slums continue to grow. This growth strains the available, mostly limited resources, infrastructure, health care systems and WaSH services (Chikozho et al. 2019; Iddi et al. 2021). Since all three WaSH services were observed from the same household at the same time, they are likely to be correlated. Therefore, it is important to explore how they correlate in ways beyond those explained by the demographic and socio-economic factors.

Our results showed some differences between the univariate and multivariate outcome models regarding estimated effect sizes. The multivariate outcome model also allowed us to estimate household and year-level correlation between the services. In particular, households with garbage disposal services were *less* likely to have improved toilet and water services, while the latter two were correlated with each other across households. Further, across the years, water and toilet services were likely to move in the same direction (improve), but not garbage disposal services.

Overall, we found that households were likelier to have improved water services than the other WaSH services. This could be attributed to many governments' and stakeholders' interventions to improve access to clean and safe water in informal urban settlements (Chikozho et al. 2019). Concerning the service status in the previous year, households with unimproved services in the previous year were more likely to have improved water service and garbage disposal in the current year. This was not the case for toilet services. This result is consistent with the findings in Iddi et al. 2021 which looked at the transition probabilities of WaSH indicators across years using a multi-state modeling approach. We found that households with older head of household are less likely to have improved services across the three services, possibly because older head of household are less likely to engage in extra income-generating activities to improve WaSH services within the households. Our study showed that households from Korogocho were likelier to have improved WaSH services than those from Viwandani, possibly due to the recent intervention programs by Umande Trust and GOAL Ireland carried out in Korogocho (Iddi et al. 2021).

5.4.1 Toilet services

An average household was less likely to have improved toilet services than other services. Households with unimproved toilet service in the previous year had a very low probability of improving it in the current year. This could be because most of these households are overcrowded and have limited spaces to construct toilet facilities, have large families, and, in most cases, have poor sanitation practices like open defecation (Simiyu 2015; Mberu et al. 2017; Iddi et al. 2021). Other than household expenditure and income, all other quantitative factors (high dwelling index, large households, recent year of interview, high ownership index, and high household rating) were associated with a higher probability of improved toilet services. The improvement across time could be due to various WaSH interventions implemented in the two slums, especially over the years. For instance, in 2006, World Toilet Association, Umande Trust, and APHRC implemented the construction of toilet projects in the two slums (Chikozho et al. 2019).

Further, our study revealed that male-headed households had a higher probability of improved toilet services than their female counterparts, similar to the findings of Iddi et al. 2021. Households with more food were less likely to have improved toilet services, in contrast to the conclusions of Iddi et al. 2021. Whereas our study used the amount of food consumed to measure food security, Iddi et al. 2021 constructed a composite score for food security.

5.4.2 Water services

Households were more likely to have improved water services than other services, on average. Households with unimproved water service in the previous year had a very high probability of improving it in the current year. This could be due to an increase in the provision of clean and safe water by the county government and other stakeholders (Chikozho et al. 2019). Generally, a lower probability of water service was associated with higher values of most quantitative factors (dwelling index, household expenditure, household income, year of interview, index for items owned within the household). Most importantly, our results showed a continuous decline in the probability of improved water services, suggesting that increased provision of improved water services did not match growing slum populations. Chikozho et al. 2019 arrived at a similar conclusion. On the other hand, larger households were more likely to have improved water services.

Rich households were associated with a high probability of improved water services, most likely because households rated as rich have extra income to spare for improving water services. Our study did not show clear differences in the probability of improved water services between the male and female-headed households. Iddi et al. 2021 made similar observations. Households that owned the houses they were residing in were more likely to have improved water services. These households probably had better structures and were more likely to invest in improved water services such as piped and taped water.

5.4.3 Garbage disposal services

About half of the surveys reflected improved garbage disposal. This relatively low value is likely due to a combination of low household income, and lack of both public garbage collection and clear garbage disposal policies.

Other than household size, ownership index (away from the place of residence), and household rating, all other quantitative factors (high dwelling index, household expenditure, household income, recent year of interview, and high ownership index) were associated with a high probability of improved garbage disposal services. The improvement through time could be due to various WaSH intervention programs previously mentioned.

Households that rented the residence were likelier to have improved garbage disposal services than owned households. In slums, garbage services fees are mandatory in most rented houses and are paid as part of the monthly rent. This observation is consistent with the findings from other studies (Banga et al. 2011; Iddi et al. 2021).

5.5 Conclusion

The results from this study provide an overview of predictors of access to WaSH services in Nairobi's slum areas of Viwandani and Korogocho. Importantly, our results point to the need for researchers and policymakers to consider the possibility of correlation between WaSH outcomes beyond those explained by the explanatory variables – WaSH outcomes are correlated at either household or year level. The evidence generated from this study could inform WaSH promoters on which particular WaSH indicator to target in order to improve all three WaSH indicators. For instance, WaSH agencies could focus more on underlying poverty-related, demographic, and socio-economic factors that continue to hinder opportunities for improved WaSH indicators.

We are hopeful that the results from this study can be used to inform the agenda of policymakers and public health practitioners who grapple continuously with the challenges faced in accessing WaSH services in Nairobi's low-income residential areas. It will also directly contribute to the growing knowledge on access to improved WaSH services in the context of slum areas in low- and middle-income countries.

Chapter 6

Conclusion

The boundary between computational inference and prediction in public health is hard to define – many techniques from statistical methods (SM) and machine learning (ML) may, in principle, be used for both perspectives. Some techniques fall entirely into one or the other domain, but many span both (Bzdok et al. 2018). In general, SM requires us to build models that incorporate our knowledge of the data-generation process or the system – and the justification typically relies on our confidence about how well we think it captures the process. On the other hand, ML methods build more flexible predictive algorithms by mainly relying on what we have observed – and the choice often rests on past performance measures in related or similar problems. The approaches are complementary for drawing conclusions about public health problems.

6.1 Key Findings

In Chapter 2, I demonstrated how the penalized Cox proportional hazard models could be extended to handle time-dependent covariates and provided R software package **pcoxtime** implementation. To solve the optimization problem, I exploited proximal gradient descent, which is intrinsically slower than other candidate methods such as those based on coordinate descent. However, until recently, only **pcoxtime** provided the ability to handle time-dependent covariates in penalized Cox proportional hazard models. Furthermore, even with the improvement of the other existing methods to handle this kind of data, **pcoxtime** has shown better convergence properties and offers more flexibility and ease in terms of model formulation and post-model predictions.

In Chapter 3, I performed two classes of analysis: first, based on yearly-cohort time-invariant datasets; second, based on fully time-varying covariates datasets. In both cases, I compared traditional Cox proportional hazard models and the hazard-based machine learning models. I found that machine learning-based methods can provide more accurate alternatives to traditional hazard-based methods in both cases. In particular, the Cox model with gradient boosting machine had the highest predictive performance score in all the comparisons done, indicating that it may be a good choice in general for problems of this nature. I also found that time-varying covariates greatly improve model prediction in machine learning.

In Chapter 4, I discussed two different ways of generating and interpreting "outcome plots" – as prediction plots focused on uncertainty in predictions, or as effect plots focused on the uncertaint of effects of a given predictor. I also compared and implemented two approaches for generating outcome plots for generalized linear models – mean-based and observed-value-based. I showed that the two approaches can produce substantially different results and that the observed-value-based approach can produce estimates that are more consistent with the observed values. The results I present showed that outcome plots provide a useful way to summarize the results of (generalized) linear (mixed) models. I implemented R software package **vareffects** which offers researchers a flexible way to generate outcome plots based on the observed-value approach, which is not currently implemented in other packages.

Lastly, in Chapter 5, I compared multivariate outcome models for binary outcomes to investigate the contribution of demographic and economic factors to water, sanitation, and hygiene (WaSH) indicators in two informal urban settlements in Kenya, Korogocho and Viwandani. The results provide an overview of the extent to which residents of these settlements can access and the status of WaSH services. Most importantly, it points to the need for researchers and policymakers to consider the possibility of correlation between WaSH outcomes beyond those explained by the explanatory variables – WaSH outcomes are correlated at either household or year level.

6.2 Future research possibilities

I briefly discuss some future research possibilities along the line of this work.

- In Chapter 2, I plan to improve the functionality of **pcoxtime**. In particular, I plan to implement a coordinate descent algorithm in place of the current proximal gradient descent approach, which should greatly improve its speed.
- In Chapter 3, I highlight computational challenges in training the neural network and random forest algorithms. It would be interesting to explore how neural network and random forest algorithms trained on the entire dataset compare to the gradient boosting model. I could consider extending the implementation of neural network and random forest to be able, during model training, to periodically save a subset of the neurons and forests, respectively, into the computer's physical hard disk instead of the RAM and only keep (or access) key parameters and components needed for various steps in the memory.

Bibliography

Ahmad, F., N. Almuayqil, S., Humayun, M., Naseem, S., Ahmad Khan, W., and Junaid, K. (2021). Prediction of COVID-19 Cases using Machine Learning for Effective Public Health Management. *Computers, Materials & Continua* 66(3), 2265–2282.

Allison, P. D. (2010). *Survival analysis using SAS: a practical guide.* 2. ed. Cary, NC: SAS Press. ISBN: 9781599946405.

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (July 2017). The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ* 5, e3544.

Andersen, P. K. and Gill, R. D. (Dec. 1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* 10(4).

Austin, P. C., Latouche, A., and Fine, J. P. (2020). A Review of the use of Time-varying Covariates in the Fine-Gray Subdistribution Hazard Competing Risk Regression Model. *Statistics in Medicine* 39(2), 103–113.

Banga, M., Lokina, R. B., and Mkenda, A. F. (Dec. 2011). Households' Willingness to Pay for Improved Solid Waste Collection Services in Kampala City, Uganda. *The Journal of Environment & Development* 20(4), 428–448.

Barzilai, J. and Borwein, J. M. (1988). Two-point Step Size Gradient Methods. *IMA Journal of Numerical Analysis* 8(1), 141–148.

Beguy, D., Elung'ata, P., Mberu, B., Oduor, C., Wamukoya, M., Nganyi, B., and Ezeh, A. (2015). Health & demographic surveillance system profile: the Nairobi urban health and demographic surveillance system (NUHDSS). *International journal of epidemiology* 44(2), 462–471.

Bender, A., Groll, A., and Scheipl, F. (2018). A Generalized Additive Model Approach to Time-to-event Analysis. *Statistical Modelling* 18(3-4), 299–321.

Berry, W. D., Golder, M., and Milton, D. (July 2012). Improving Tests of Theories Positing Interaction. *The Journal of Politics* 74(3), 653–671.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (Jan. 2011). A review of survival trees. *Statistics Surveys* 5(none).

Brambor, T., Clark, W. R., and Golder, M. (2006). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis* 14(1), 63–82.

Breslow, N. E. (1972). Contribution to Discussion of Paper by Dr Cox. *J. Roy. Statist. Soc., Ser. B* 34, 216–217.

Bzdok, D., Altman, N., and Krzywinski, M. (Apr. 2018). Statistics versus machine learning. *Nature Methods* 15(4), 233–234.

Bzdok, D., Engemann, D., and Thirion, B. (Nov. 2020). Inference and Prediction Diverge in Biomedicine. *Patterns* 1(8), 100119.

Bzdok, D. and Ioannidis, J. P. (Apr. 2019). Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences* 42(4), 251–262.

Cagan, R. and Meyer, P. (2017a). *Rethinking Cancer: Current Challenges and Opportunities in Cancer Research.*

Cagan, R. and Meyer, P. (Apr. 2017b). Rethinking cancer: current challenges and opportunities in cancer research. *Disease Models & Mechanisms* 10(4), 349–352.

Cheon, S., Agarwal, A., Popovic, M., Milakovic, M., Lam, M., Fu, W., DiGiovanni, J., Lam, H., Lechner, B., Pulenzas, N., Chow, R., and Chow, E. (Jan. 2016). The accuracy of clinicians' predictions of survival in advanced cancer: a review. *Annals of Palliative Medicine* 5(1), 229–229.

Chib, S. (June 1998). Analysis of multivariate probit models. *Biometrika* 85(2), 347–361.

Chikozho, C., Kadengye, D. T., Wamukoya, M., and Orindi, B. O. (2019). Leaving no one behind? Analysis of trends in access to water and sanitation services in the slum areas of Nairobi, 2003–2015. *Journal of Water, Sanitation and Hygiene for Development* 9(3), 549–558.

Chow, E., Harth, T., Hruby, G., Finkelstein, J., Wu, J., and Danjoux, C. (June 2001). How Accurate are Physicians' Clinical Predictions of Survival and the Available Prognostic Tools in Estimating Survival Times in Terminally III Cancer Patients? A Systematic Review. *Clinical Oncology* 13(3), 209–218.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society B (Methodological)* 34(2), 187–202.

Cygu, S., Dushoff, J., and Bolker, B. M. (June 2021). pcoxtime: Penalized Cox Proportional Hazard Model for Time-dependent Covariates. *arXiv:2102.02297 [stat]*. arXiv: 2102.02297.

D. Redelings, M., Sorvillo, F., V. Smith, L., and Greenland, S. (Oct. 2012). Why Confidence Intervals Should be Used in Reporting Studies of Complete Populations. *The Open Public Health Journal* 5(1).

Dai, B. and Breheny, P. (2019). Cross Validation Approaches for Penalized Cox Regression. *ArXiv Preprint ArXiv:1905.10432*.

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (Dec. 2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* 6(1), 54.

Dushoff, J., Kain, M. P., and Bolker, B. M. (June 2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution* 10(6). Ed. by R. B. O'Hara, 756–759.

Duursma, R. and Robinson, A. (2003). Bias in the mean tree model as a consequence of Jensen's inequality. *Forest Ecology and Management* 186(1-3), 373–380.

Efron, B. and Hastie, T. (2016). *Computer age statistical inference: algorithms, evidence, and data science*. Institute of Mathematical Statistics monographs. New York, NY: Cambridge University Press. ISBN: 9781107149892.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (Feb. 2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639), 115–118.

Fang, D., Sun, R., and Wilson, J. R. (Jan. 2018). Joint modeling of correlated binary outcomes: The case of contraceptive use and HIV knowledge in Bangladesh. *PLOS ONE* 13(1). Ed. by I. Puebla, e0190917.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Thousand Oaks, Calif: Sage Publications. ISBN: 9780761922797 9780761922803.

Fox, J. and Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software* 32(1), 1–24.

Fujino, Y., Suzuki, Y., Ajiki, T., Tanioka, Y., Ku, Y., and Kuroda, Y. (Feb. 2003). Predicting factors for survival of patients with unresectable pancreatic cancer: a management guideline. *Hepato-Gastroenterology* 50(49), 250–253.

Goeman, J. J. (2010). L1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal* 52(1), 70–84.

Goeman, J. J. (2018). **Penalized** R Package. R package version 0.9-51.

Gorst-Rasmussen, A. and Scheike, T. H. (2012). Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model. *Journal of Statistical Software* 47(1), 1–17.

Gui, J. and Li, H. (2005). Penalized Cox Regression Analysis in the High-dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics* 21(13), 3001–3008.

Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R. L., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., Matheson, L., Ashley, D. M., and Venkatesh, S. (Mar. 2014). Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open* 4(3), e004007.

Hanmer, M. J. and Ozan Kalkan, K. (2013). Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science* 57(1), 263–277.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (Jan. 2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25(1), 65–69.

Harrell Jr, F. E. (2015). *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer-Verlang.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* 15(4), 361–387.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* Springer Science & Business Media.

Hayward, J., Alvarez, S. A., Ruiz, C., Sullivan, M., Tseng, J., and Whalen, G. (July 2010). Machine learning of clinical performance in a pancreatic cancer database. *Artificial Intelligence in Medicine* 49(3), 187–195.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (June 2000). Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* 56(2), 337–344.

Heagerty, P. J. and Paramita Saha-Chaudhuri, packaging by (2012). *risksetROC: Riskset ROC curve estimation from censored survival data.* R package version 1.0.4.

Heagerty, P. J. and Zheng, Y. (Mar. 2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics* 61(1), 92–105.

Hinton, G. E. and Salakhutdinov, R. R. (July 2006). Reducing the Dimensionality of Data with Neural Networks. *Science* 313(5786), 504–507.

Hochstein, A., Ahn, H.-I., Leung, Y. T., and Denesuk, M. (2013). Survival Analysis for HDLSS Data with Time Dependent Variables: Lessons from Predictive Maintenance at a Mining Service Provider. In: *Proceedings of 2013 IEEE International Conference on Service Operations and Logistics, and Informatics.* IEEE, 372–381.

Iddi, S., Akeyo, D., Bagayoko, M., Kiwuwa-Muyingo, S., Chikozho, C., and Kadengye, D. T. (Nov. 2021). Determinants of transitions in water and sanitation services in two urban slums of Nairobi: A multi-state modeling approach. *Global Epidemiology* 3, 100050.

Ioannidis, J. P. A. (Apr. 2018). The Proposal to Lower *P* Value Thresholds to .005. *JAMA* 319(14), 1429.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (Oct. 2014). Random survival forests for competing risks. *Biostatistics* 15(4), 757–773.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (Sept. 2008). Random survival forests. *The Annals of Applied Statistics* 2(3).

Ivanova, A., Molenberghs, G., and Verbeke, G. (2016). Mixed models approaches for joint modeling of different types of responses. *Journal of Biopharmaceutical Statistics* 26(4), 601–618.

Jordan, M. I. and Mitchell, T. M. (July 2015). Machine learning: Trends, perspectives, and prospects. *Science* 349(6245), 255–260.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (Dec. 2018a). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology* 18(1), 24.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (Feb. 2018b). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology* 18(1), 24.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13, 8–17.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research* 20(129), 1–30.

LaLonde, A., Love, T., Thurston, S. W., and Davidson, P. W. (Nov. 2019). Discovering structure in multiple outcomes models for tests of childhood neurodevelopment. *Biometrics.*

LeCun, Y., Bengio, Y., and Hinton, G. (May 2015). Deep learning. *Nature* 521(7553), 436–444.

Leeper, T. J. (2017). Interpreting regression results using average marginal effects with R's margins. *Reference manual* 32.

Leeper, T. J., Arnold, J., Arel-Bundock, V., and Leeper, M. T. J. (2017). Package 'margins'. *accessed December* 5, 2019.

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.7.3.

Marathe, M. V. and Ramakrishnan, N. (July 2013). Recent Advances in Computational Epidemiology. *IEEE Intelligent Systems* 28(4), 96–101.

Mberu, B., Béguy, D., and Ezeh, A. C. (2017). Internal Migration, Urbanization and Slums in Sub-Saharan Africa. In: *Africa's Population: In Search of a Demographic Dividend.* Ed. by H. Groth and J. F. May. Cham: Springer International Publishing, 315–332.

Mhasawade, V., Zhao, Y., and Chunara, R. (2020). Machine Learning in Population and Public Health.

Mihaylov, I., Nisheva, M., and Vassilev, D. (Mar. 2019). Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies. *Information* 10(3), 93.

Miro-Quesada, G., Del Castillo, E., and Peterson, J. J. (2004). A Bayesian approach for multiple response surface optimization in the presence of noise variables. *Journal of Applied Statistics* 31(3), 251–270.

Montazeri, M., Montazeri, M., Montazeri, M., and Beigzadeh, A. (Jan. 2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care* 24(1), 31–42.

O'Brien, S. M. and Dunson, D. B. (Sept. 2004). Bayesian Multivariate Logistic Regression. *Biometrics* 60(3), 739–746.

Papachristou, N., Puschmann, D., Barnaghi, P., Cooper, B., Hu, X., Maguire, R., Apostolidis, K., P. Conley, Y., Hammer, M., Katsaragakis, S., M. Kober, K., D. Levine, J., McCann, L., Patiraki, E., P. Furlong, E., A. Fox, P., M. Paul, S., Ream, E., Wright, F., and Miaskowski, C. (Dec. 2018). Learning from data to predict future symptoms of oncology patients. *PLOS ONE* 13(12). Ed. by W. Mumtaz, e0208808.

Parikh, N. and Boyd, S. (2014). Proximal Algorithms. *Foundations and Trends in Optimization* 1(3), 127–239.

Park, M. Y. and Hastie, T. (2007). L1-regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society B (Statistical Methodology)* 69(4), 659–677.

Paulus, M. P. and Thompson, W. K. (May 2021). Computational approaches and machine learning for individual-level treatment predictions. *Psychopharmacology* 238(5), 1231–1239.

Pierce, G. (2017). Why is basic service access worse in slums? A synthesis of obstacles. *Development in Practice* 27(3), 288–300.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. (Mar. 2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* 2(3), 158–164.

Santos, B. S. dos, Steiner, M. T. A., Fenerich, A. T., and Lima, R. H. P. (Dec. 2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering* 138, 106120.

Schaik, P. van, Peng, Y., Ojelabi, A., and Ling, J. (July 2019). Explainable statistical learning in public health for policy development: the case of real-world suicide data. *BMC Medical Research Methodology* 19(1), 152.

Seow, H., Barbera, L., Sutradhar, R., Howell, D., Dudgeon, D., Atzema, C., Liu, Y., Husain, A., Sussman, J., and Earle, C. (Mar. 2011). Trajectory of Performance Status and Symptom Scores for Patients With Cancer During the Last Six Months of Life. *Journal of Clinical Oncology* 29(9), 1151–1158.

Seow, H., Dhaliwal, G., Fassbender, K., Rangrej, J., Brazil, K., and Fainsinger, R. (Jan. 2016). The Effect of Community-Based Specialist Palliative Care Teams on Place of Care. *Journal of Palliative Medicine* 19(1), 16–21.

Seow, H., Tanuseputro, P., Barbera, L., Earle, C., Guthrie, D., Isenberg, S., Juergens, R., Myers, J., Brouwers, M., and Sutradhar, R. (Apr. 2020). Development and Validation of a Prognostic Survival Model With Patient-Reported Outcomes for Patients With Cancer. *JAMA Network Open* 3(4), e201768.

Shi, C.-F., Li, M., and Dushoff, J. (Apr. 2017). Evidence that promotion of male circumcision did not lead to sexual risk compensation in prioritized Sub-Saharan countries. *PLOS ONE* 12(4). Ed. by M. A. Price, e0175928.

Simiyu, S. (Oct. 2015). Socio-economic dynamics in slums and implications for sanitation sustainability in Kisumu, Kenya. *Development in Practice* 25(7), 986–996.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 39(5), 1.

Sohn, I., Kim, J., Jung, S.-H., and Park, C. (2009). Gradient Lasso for Cox Proportional Hazards Model. *Bioinformatics* 25(14), 1775–1781.

Sorlie, T. and Tibshirani, R. (2003). Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets. *Proc Natl Acad Sci USA* 100, 8418–8423.

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., and Brodaty, H. (Nov. 2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports* 10(1), 20410.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15(56), 1929–1958.

Sylvestre, M.-P. and Abrahamowicz, M. (2008). Comparison of Algorithms to Generate Event Times Conditional on Time-dependent Covariates. *Statistics in Medicine* 27(14), 2618–2634.

Szucs, D. and Ioannidis, J. P. A. (Aug. 2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience* 11, 390.

Therneau, T., Crowson, C., and Atkinson, E. (2017). Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. *Survival Vignettes.*

Therneau, T. M. (2020). *A Package for Survival Analysis in R.* R package version 3.2-7.

Therneau, T. M. (2022). *A Package for Survival Analysis in R.* R package version 3.3-1.

Thomas, L. and Reyes, E. M. (2014). Tutorial: Survival Estimation for Cox Regression Models with Time-Varying Coefficients Using *SAS* and *R. Journal of Statistical Software* 61(Code Snippet 1).

Tibshirani, R. J. (Jan. 2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* 7(none).

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (July 2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* 101(1), 5–22.

Wang, P., Li, Y., and Reddy, C. K. (Nov. 2019). Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys* 51(6), 1–36.

Wasserstein, R. L. and Lazar, N. A. (Apr. 2016). The ASA Statement on p -Values: Context, Process, and Purpose. *The American Statistician* 70(2), 129–133.

Wiemken, T. L. and Kelley, R. R. (Apr. 2020). Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health* 41(1), 21–36.

Yang, Y. and Zou, H. (2013). A Cocktail Algorithm for Solving the Elastic Net Penalized Cox's Regression in High Himensions. *Statistics and its Interface* 6(2), 167–173.

Yao, W., Frydman, H., Larocque, D., and Simonoff, J. S. (Aug. 2021). Ensemble Methods for Survival Data with Time-Varying Covariates. *arXiv:2006.00567 [stat]*. arXiv: 2006.00567.

Yu, W., Wardrop, N. A., Bain, R. E. S., Lin, Y., Zhang, C., and Wright, J. A. (Mar. 2016). A Global Perspective on Drinking-Water and Sanitation Classification: An Evaluation of Census Content. *PLOS ONE* 11(3), 1–17.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society B (Statistical Methodology)* 67(2), 301–320.