Estimating Mortality Risk in Populations and Individuals: Applications of Bayesian and Machine Learning Methods

Reed Sorensen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Christopher Murray, Chair

Ashkan Afshin

Aleksandr Aravkin

Abraham Flaxman

Program Authorized to Offer Degree:
Health Metrics Sciences

University of Washington

Abstract

Estimating Mortality Risk in Populations and Individuals:
Applications of Bayesian and Machine Learning Methods

Reed Sorensen

Chair of the Supervisory Committee:
Christopher Murray
Department of Health Metrics Sciences

Understanding mortality risk, including its distribution and determinants, is fundamental to the health sciences and any effort to prevent needless deaths. This dissertation is an exploration of modern methods for estimating mortality risk in populations and individuals. While Bayesian statistics and machine learning have both benefited from advances in modern computing, they are polar opposites in terms of modeling strategy. Bayesian methods are rigid by design – rigid in a particular way determined by the modeler – in an effort to provide contextual information to the model. Machine learning methods are more flexible, seeking the information content of the data wherever it may exist. Chapter 1 of this dissertation describes a novel Bayesian method called a *spline cascade* that is capable of characterizing how a non-linear curve varies across hierarchical subsets of a dataset. We developed the method to model age patterns of COVID-19 mortality for global locations. Chapter 2 demonstrates how machine learning and variable attribution methods can and should be used in analytic epidemiology. We used XGBoost and SHAP values to investigate patterns in the relationship between anthropometric measurements and mortality risk, adjusting for age. Chapter 3 is a simulation study comparing Bayesian spline cascades, XGBoost and existing methods for estimating child mortality risk in 193 countries. We develop a theory of model validation and discuss the role of Bayesian statistics and machine learning in the health sciences.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Ashkan, Sasha, Abie and Chris for guiding me through the long and rewarding process of completing this research. It has been a privilege to have each of them as a direct supervisor at different points in the last six years. Thank you to the countless people at the University of Washington who supported this work, especially Haidong Wang and David Pigott who led the data gathering efforts for the under-5 mortality and COVID-19 mortality models, respectively, and Peng Zheng who developed MRTool. Thank you to Emm Gakidou and Alan Lopez for serving on my committee before the topic shifted away from the health effects of smoking, and to Michael LeBlanc for agreeing to be my Graduate Student Representative. Finally, thank you to my friends and family who celebrated the milestones with me along the way.

DEDICATION

to Hillary

Chapter 1

THE BAYESIAN SPLINE CASCADE METHOD FOR ESTIMATING AGE PATTERNS OF COVID-19 MORTALITY

1.1 Introduction

1.1.1 Overview

In this chapter, we introduce a novel method called a *Bayesian spline cascade* for estimating non-linear effects in a nested hierarchy. The method was developed to estimate age patterns of COVID-19 mortality in every country by forming a nested hierarchy of locations. We first describe the importance of modeling COVID-19 age patterns and summarize how others have attempted to address common methodological challenges. We then briefly describe the Bayesian spline cascade method and its novel features. Then, we look into the wider literature on modeling non-linear patterns in order to compare and contrast Bayesian spline cascades with other methods. The rest of the chapter describes Bayesian spline cascades in detail, both mathematically and by example, and discusses how this method is used in practice as part of a larger COVID-19 forecasting project. The purpose of the present study is to demonstrate the Bayesian spline cascade method; quantitative results of the model are presented in separate manuscript currently under review for publication in The Lancet (title: "Variation in the COVID-19 infection-fatality ratio by age, time and geography during the pre-vaccine era").

1.1.2 COVID-19 mortality age patterns

Characterizing the age pattern of Coronavirus disease-2019 (COVID-19) mortality is a key step in understanding the pandemic and its impact on health systems. Mortality serves as the most conspicuous sign of COVID-19 transmission and is the clinical outcome of greatest

concern. We parameterize mortality as the number of cumulative deaths divided by the size of the population; the term "age pattern" refers to how this parameter varies as a function of age. As of August 2021, more than four million deaths due to COVID-19 have been reported globally [1]. Because older individuals are at higher risk of dying once infected with COVID-19 [2], attempts to explain geographic variation in death rates must account for a population's age distribution and its age-specific mortality risk. Tracking how mortality age patterns change through time provides insights into patterns of community transmission, which in turn can guide intervention strategies targeted to the local population. The age pattern of mortality is also a critical input for calculating age-specific levels of the infection-fatality ratio (IFR), a metric of clinical importance that quantifies the likelihood of dying among those infected with the SARS-CoV-2 virus.

Previous efforts to characterize age patterns of COVID-19 mortality have varied in their approaches. To address inconsistent categorization of age-disaggregated death data, harmonization strategies often involve making simplifying assumptions that decrease the information content of the data. For example, some studies use only the age delineations that are common between datasets: "[...] all countries report it differently, with under- and over-65 as the only age group with comparable data. Thus, we created a dataset with age-specific mortality for under- and over 65 for all 5 countries" [3]. Others treat non-identical age groups as identical: "As the reported age-groups varied by country, the age group with an upper bound of 59 was chosen as the reference group where possible. Where this was not an available age-group, the age-group with an upper bound of 64 was selected as the reference" [2]. The most common method is to use the midpoint of the age group as an approximation. The midpoint approach may be suitable for smaller age groups but becomes increasingly inaccurate for non-linear patterns as the coverage of the age group increases. In all cases, special consideration must be given to the open-ended age group (e.g. age 80 and above), the upper bound of which can be arbitrary or undefined. The most rigorous attempts to address this issue use information about the age distribution of the population within the group, for example, using the median of the age distribution as the group's midpoint [4] or

using indirect standardization [5].

1.1.3 The Bayesian spline cascade method

In this article, we present a novel modeling approach called a *Bayesian spline cascade*. The spline cascade method was originally developed to model location-specific age patterns of COVID-19 mortality but applies generally to the estimation of non-linear patterns in a nested hierarchy. Briefly, the method begins by fitting a mixed effects meta-regression using all data to inform the shape of the age curve, then passes the estimated spline coefficients as priors to models fit on location-specific subsets of the data. The process may be repeated an arbitrary number of times to form a hierarchical cascade structure. The effect is that local age pattern predictions, where data may be sparse, are augmented by global and regional information about the shape of the age curve. The method takes advantage of a unique property of meta-regression models, models that incorporate measurement error as part of the dependent variable. In locations where data are less informative (high measurement error), the shape of the age curve is influenced more by the prior from the parent model. In locations where data are more informative (low measurement error), the data have relatively more influence. An additional benefit is that the underlying optimization package `MRTool` [6; 7] employs a method for fitting a continuous, non-linear curve from data reported as age groups, thus avoiding the need to make simplifying assumptions like using the midpoint of an age group as the observed value. Apart from providing a coherent solution to the challenge of COVID-19 data synthesis, the primary benefit of the method is that it provides greater local specificity in age patterns. It fills in information gaps created by coarse age groups and suppressed data, and is capable of making out-of-sample predictions. The method is used to model several parameters (e.g. deaths, cases, hospitalizations, time trends) as part of IHME's ongoing COVID-19 forecasting project.

1.1.4 Comparisons to other models

The primary antecedent of Bayesian spline cascades is DisMod, a disease modeling framework developed by Flaxman and colleagues [8]. DisMod uses information about multiple disease parameters (e.g. prevalence, incidence rate, remission rate, mortality rate) and the mathematical relationships between them to make location-specific estimates that are internally consistent. To model the age pattern, DisMod estimates a global age effect with a spline and passes the posterior predictive distribution as a prior to region-specific models. While similar in aim, the present framework differs in structure and scope. It derives empirical priors from estimated coefficients rather than the predictive distribution of a curve. It is also implemented as a generic meta-regression package and not embedded in a wider integrative disease modeling system. Finally, the underlying Bayesian meta-regression package `MRTool` [6] comes equipped with additional functionality for outlier trimming, variable selection, and a wider array of options for priors and shape constraints.

To our knowledge, using estimated spline coefficients as priors in a hierarchical cascade has not been discussed previously in the literature. Spline models that go under the name "hierarchical" typically refer to the use of random effects on spline coefficients, for example, estimating a population-average curve with stochastic subject-specific variation [9; 10; 11]. One such model incorporates a Bayesian prior on the variance of random effects [12] but not the fixed effect coefficients as in the *spline cascade* framework. Spline models implementing an "empirical Bayesian" approach typically are using priors to regularize the complexity of a curve [13], or empirically determine the number and/or location of knots [14; 15]. We found a model implementing "Bayesian hierarchical splines" for descriptive epidemiology [16], but it did not make use of priors and the hierarchical element dealt only with the degree of smoothing.

1.1.5 Purpose of this study

In the following sections, we introduce spline cascades mathematically and by example. Using real-world data, we highlight instances where the unique aspects of the model allow for improved modeling of age patterns. We also discuss how this model directly contributes to ongoing research in estimating COVID-19 infection fatality ratios for global populations.

1.2 Methods

1.2.1 Overview

The Bayesian spline cascade method first fits a mixed effects meta-regression using all available data to inform the shape of the age pattern, then uses the estimated coefficients as Bayesian priors in subsequent models fit on subsets of the data. This process may be repeated for an arbitrary number of hierarchical stages, including subsets based on geography, time or other dimensions. The framework is an extension of the open source `MRTool` meta-regression package [6; 7]. `MRTool`'s non-linear estimation method, B-splines, is equipped with additional functionality that allows for a high degree of control over the functional form of the age effect. First, we develop notation for the general Bayesian cascade framework and discuss how it applies to spline cascades. Then we describe methods for estimating age patterns of COVID-19 mortality in a cascade that includes stages for global, region-specific and location-specific models.

1.2.2 Technical specifications for hierarchical Bayesian cascades

Spline cascades are part of a more general class of models we call hierarchical Bayesian cascades. The cascade framework requires a dataset consisting of observations $i = 1, \ldots, N$ classified into groups $j = 1, \ldots, J$. The groups form a nested hierarchy with sequential stages $k \in \{1, 2, \ldots, K\}$. Each observation i is associated with K groups corresponding to K stages of the hierarchy. Figure 1.1 shows an example cascade with three stages and 10 groups. Groups in the same level are assumed to be mutually exclusive.

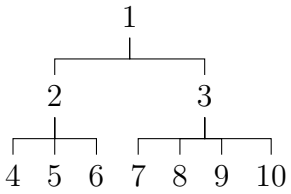

Figure 1.1: Example of a three-stage cascade structure with 10 component groups

We use $\mathcal{J}_k$ to denote which groups j are in level k. In Figure 1.1, for example, level-specific groups are $\mathcal{J}_1 = \{1\}$, $\mathcal{J}_2 = \{2, 3\}$, $\mathcal{J}_3 = \{4, 5, 6, 7, 8, 9, 10\}$. To describe the nested nature of the cascade, the $\uparrow$ operator signifies which groups are descendants of another group. For example, $4_\uparrow = 5_\uparrow = 6_\uparrow = \{2\}$ indicates that groups 4, 5 and 6 are descendants of group 2. Note that the group at the top of the hierarchy has no ancestor by definition, i.e. $1_\uparrow = \{\}$.

The modeling cascade proceeds sequentially from level to level, beginning with $k = 1$ which is constituted by one group. For each $j \in \mathcal{J}_k$, we fit a model using only observations in group j:

$$y_{ij}^k = (x_{ij}^k)^T \beta_j^k + \epsilon_{ij}^k, \quad \epsilon_{ij}^k \sim N(0, (\sigma_j^k)^2) \tag{1.1}$$

The use of k in the superscript denotes the level of the hierarchy. Estimates of β_j^k are informed by the estimated β^{k-1} from the parent model in the form of Bayesian priors:

$$\beta_j^k = \hat{\beta}_{j_\uparrow}^{k-1} + \nu_j^k, \quad \nu_j^k \sim N(0, \lambda^k V_{j_\uparrow}^{k-1}). \tag{1.2}$$

In the case of $\mathcal{J}_1$ where no parent model is available, priors may be user-defined or left as non-informative. Users may modify the strength of the prior passed to all models in a stage k by setting λ^k. For Gaussian priors, λ influences the strength of the prior as a multiplier of the prior's standard deviation. Higher values of λ^k give relatively more weight to the data and lower values of λ^k give more weight to the prior. Finally, apart from the priors on β that are passed down the cascade, β in all stages is constrained to lie in set Ω, specified in the first-stage model through equality and inequality constraints.

To summarize in terms of the negative log-likelihood, we fit the following model for each

group j:

$$\hat{\beta}_j^k = \arg\min_{\beta \in \Omega} \sum_{i=1}^{N_j} g(\beta; y_{ij}^k, \Sigma_j^k) + p(\beta; \hat{\beta}_{j\uparrow}^{k-1}, \lambda^k V_{j\uparrow}^{k-1}) \tag{1.3}$$

The estimated betas $\hat{\beta}$ for a model j in stage k are optimized, subject to constraints Ω, by minimizing a loss function g with priors p that come from the parent model one stage higher in the hierarchy. Posterior uncertainty distributions are obtained with a modified form of the parametric bootstrap; see Zheng and colleagues for details [7].

1.2.3 Spline cascade definition and implementation

Spline cascades are a type of hierarchical Bayesian cascade designed to capture variation in non-linear patterns among subgroups in the data. The rationale is that the non-linear relationship between a predictor and the outcome (e.g. the relationship between age and mortality) is not independent among subgroups of the data (e.g. various locations). Fitting splines separately for each subgroup would neglect useful information about their similarities. By structuring the subgroups as a hierarchy, information about the shape of the non-linear relationship can be shared in the form of Bayesian priors. Non-linear estimation methods have a tendency to overfit in data-sparse regions of the feature space [17]. Splines in particular are known for having unstable behavior at the extremes of the data [18]. With spline cascades, this instability is mitigated by using information from models higher in the hierarchy in which data sparsity is less of a concern. When the model in a cascade is a meta-regression, which includes consideration for measurement error in the dependent variable, spline cascades have the desirable property of relying more heavily on the prior when measurement error is high and more heavily on the data when measurement error is low.

Our implementation of the spline cascade method begins by fitting a single mixed effects meta-regression on all available data using `MRTool` [6; 7], including one B-spline and an arbitrary number of other terms. The coefficients estimated for the spline, but not other terms, are passed on as priors to second-stage models fit on subsets of the data. This has the effect of informing the shape of the spline (i.e. the relative age pattern) in second-

stage models while allowing the other parameters to be re-estimated. This process may be repeated for an arbitrary number of stages. If desired, users can fix as constant the non-spline coefficient values estimated in the first-stage model. Predictions are made with the lowest available model in the hierarchy. A benefit of the cascade structure, as compared to a model with random effects on each coefficient of the spline, is that models are re-fit on each subset of the data. This allows for fine-grained control over model settings intended to impact only a subset of the feature space, such as trimming, priors and constraints, while continuing to share information across subsets.

`MRTool` is a Bayesian optimization package for conducting meta-regression [6; 7]. It implements mixed effects meta-regression in which group-level variability is modeled explicitly. In public health, grouped observations often occur when a study reports observations stratified by age or sex. Notably, `MRTool` is capable of taking heterogeneous age groups as an independent variable, a characteristic especially well-suited for COVID-19 mortality data. It does this by integrating over the whole span of an age group rather than making a simplifying assumption like taking the mid-point. `MRTool` utilizes B-splines, or basis splines, which allow a non-linear curve to be expressed as a linear combination as in linear regression. As Zheng and colleagues explain, "A spline basis is a set of piecewise polynomial functions with designated degree and domain. If we denote polynomial order by p, and the number of knots by k, we need $p + k$ basis elements s_j^p, which can be generated recursively as illustrated in [Figure 1.2]. Given such a basis, we can represent any dose-response relationship as the linear combination of the spline basis elements, with coefficients $\beta \in \mathbb{R}^{p+k}$:

$$f(t) = \sum_{j=1}^{p+k} \beta_j^p s_j^p(t)." \tag{1.4}$$

Splines in `MRTool` may be specified with a variety of priors and shape constraints, including monotonicity, convexity/concavity, priors on derivative values or function values, and linearity in the tail segments. This level of control is desirable when users have prior information about the shape or complexity of a non-linear effect, for example, that mortality rates should increase monotonically with age after a certain age.

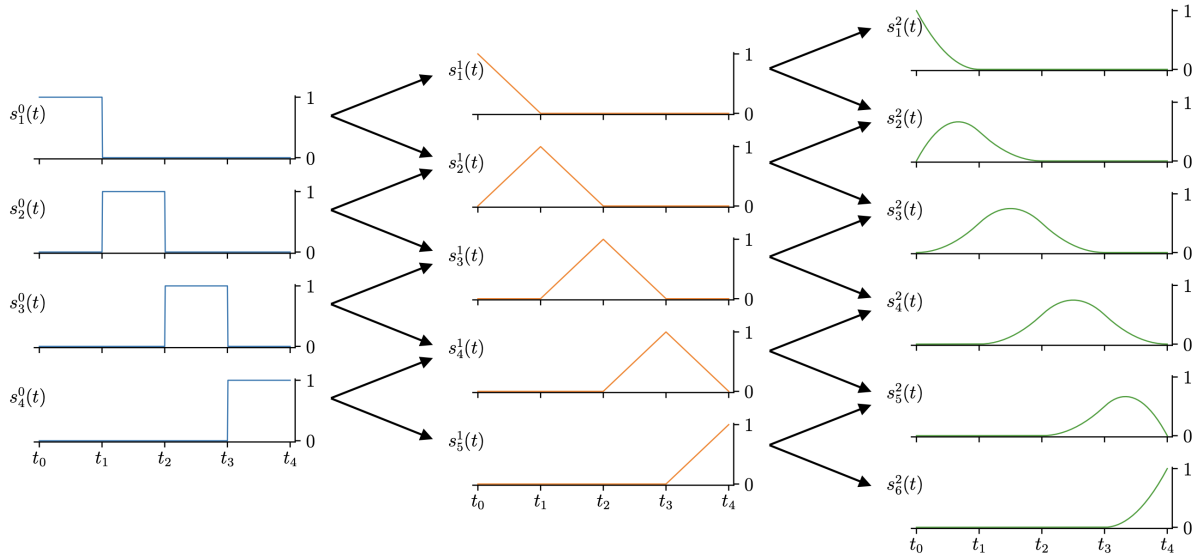

Figure 1.2: Recursive generation of B-spline basis elements (orders 0, 1, 2). Source: [7]

This framework is implemented in the R programming language. Code is available on GitHub [19]. We used the `reticulate` R package [20] to convert `MRTool` functions from Python to R.

1.2.4 Data sources, pre-processing and specifications for the example model

We obtained age-disaggregated cumulative death counts primarily from public governmental websites, such as the National Center for Health Statistics (NCHS) [21]. A list of sources is available upon request. They are listed in full in another manuscript currently under review for publication (title: "Variation in the COVID-19 infection-fatality ratio by age, time and geography during the pre-vaccine era") and will be made available in the Global Health Data Exchange [22]. We used counts of cumulative deaths from the most recent observation in a location. We adjusted the upper bound of the oldest age group such that the midpoint of the age group is equal to the mean age of the population in that group. The cascade specified as a nested geographical hierarchy with stages for global, region-specific and location-specific

models. The nested hierarchy structure was a natural choice corresponding to geographic classifications used for modeling in the Global Burden of Disease study, although in principle other structures are possible. For example, observations fully representative of a parent location might be used to fit a model which is then used as a prior for a model utilizing only observations representative of sub-locations. However, it is uncommon for data to come in this form; for example, few if any studies utilize a sampling strategy intended to be representative of the global population. The dependent variable is, for an observation indexed by location and age group, cumulative deaths at time t divided by population of the age group. An observation's standard error is calculated as the square root of binomial variance. This modeling choice gives greater weight to observations from locations with larger population sizes. The dependent variable and its standard error are transformed into logit space using the Delta method. The spline for the age effect is specified with linear right tails and an increasing monotonicity constraint for ages 10 and above. The linear tails reduce flexibility in the extremes of the spline, which we deemed important because unstable behavior of the model at older ages is not desirable. We implemented the monotonicity constraint because the data showed a clear increasing trend with age, and to our knowledge there is no plausible reason that the age effect should reverse at older ages.

1.3 Results

Figure 1.3 shows mortality age patterns for the global and region-specific stages of Model 1, the model specified as a nested geographical hierarchy of cumulative death rates. For the global model and all region-specific models, the age effect forms a J shape by decreasing until approximately age 5 to 10, then increasing monotonically as age increases. The size of the points are proportional to the inverse standard error of the logit-scale death rate. Among the region-specific estimates, high-income countries have the steepest age slope at older ages. Low- and middle-income countries have a relatively flat age effect at older ages.

Figure 1.4 demonstrates that the spline cascade framework makes out-of-sample predictions for locations with suppressed or missing data. For the District of Colombia NCHS

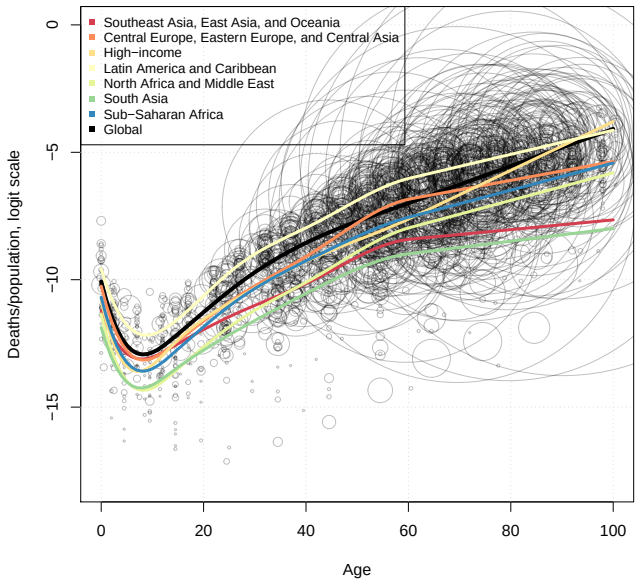

Figure 1.3: Age pattern of cumulative deaths/population, global and region-specific models

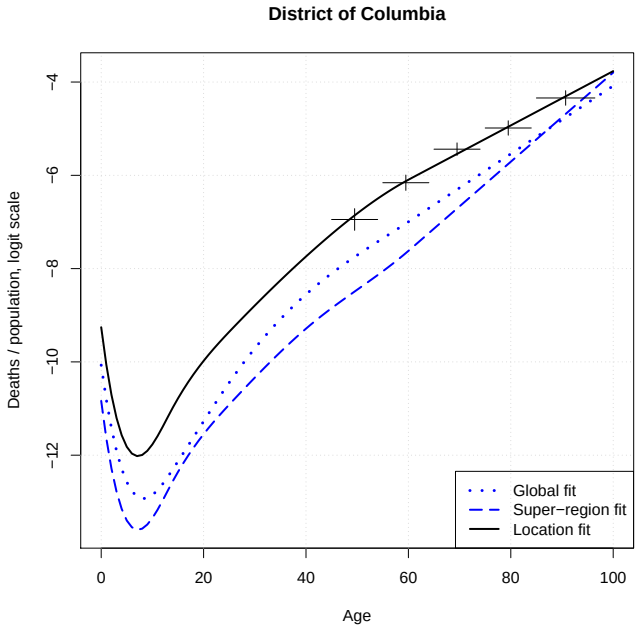

Figure 1.4: Age pattern of cumulative deaths/population, Washington D.C.

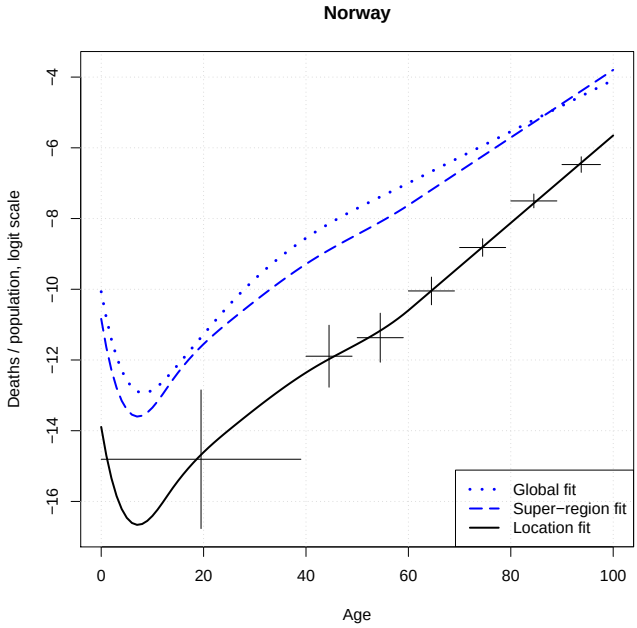

Figure 1.5: Age pattern of cumulative deaths/population, Norway

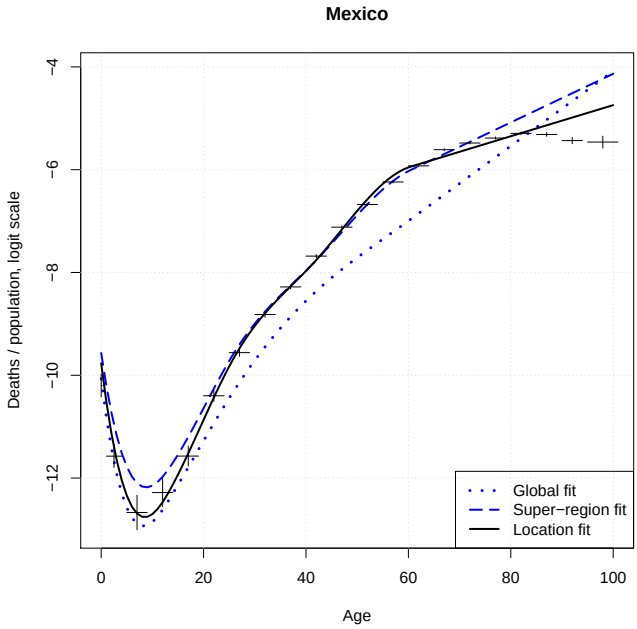

Figure 1.6: Age pattern of cumulative deaths/population, Mexico

data shown in the figure [21], counts are suppressed when there are fewer than 10 cumulative deaths. Note that the prediction follows the data where it exists at older ages, and otherwise uses the prior information from the region-specific model (blue line with thick dashes) to infer the shape of the curve at lower ages. The t-shaped symbols represent age-specific observations; the horizontal width is the coverage of the age group, and the vertical height is the size of the 95% confidence interval.

Figure 1.5 shows how the model fills in information gaps created by coarse age groups. Norway's youngest age group is 0 to 39 years of age. The model finds a continuous age pattern that is consistent with the age group's overall death rate, based on the prior from Norway's region.

Figure 1.6 is an example of where the data disagree with the prior, and the model finds a balance between them in order to make the prediction. The age pattern in Mexico's mortality data flattens and even decreases at older ages, but the prior from the region-specific model suggests that the age effect should continue increasing. The degree to which the model follows the data instead of the prior is a user-defined setting; increasing λ widens the prior probability distribution and gives greater relative weight to the data.

1.4 Discussion

In this study, we introduced the Bayesian spline cascade method conceptually, mathematically and by example. The method was originally developed as a custom solution to the challenge of synthesizing age-specific COVID-19 mortality data from global locations. Specific challenges include:

- incompatibility of age group classifications across locations;
- lack of granularity in age group classifications, i.e. the width of some age groups being too wide to describe meaningful differences;
- suppressed or otherwise missing data for certain age groups in a location;

- the need to share information about age patterns across locations and predict for out-of-sample locations.

With real-world data, we demonstrated how the Bayesian spline cascade method addresses these challenges. Given that such issues are common in descriptive epidemiology and other fields that attempt to synthesize disparate data sources, this approach to describing population-level phenomena is likely to have a wide range of applications. Indeed, the spline cascade framework has been used to model age patterns of cases, age patterns of hospitalizations and vaccination time trends in service of COVID-19 forecasting efforts. A function for running spline cascade models is part of the mrtoolr package available on GitHub (https://github.com/ihmeuw-msca/mrtoolr) [19].

The final results of the mortality age pattern model are presented in a separate manuscript currently under review for publication ("Variation in the COVID-19 infection-fatality ratio by age, time and geography during the pre-vaccine era" submitted to The Lancet). Still, it is worth discussing the general value of improvements in mortality age pattern estimation. First, because a substantial amount of variability in COVID-19 mortality rates is explained by age, age standardization is necessary for making meaningful comparisons between populations. The process of age standardization depends on information about the age distribution of a population and mortality risk conditional on age. Second, improved estimation methods can find patterns than have previously been undiscovered or underappreciated. For example, our analysis shows a clear decreasing trend among children aged 0 to 10 years, followed by a monotonically increasing trend through older ages. The J-shaped pattern of age-specific COVID-19 mortality has not been discussed widely in the literature [2]. We discuss this observation extensively in the other manuscript ("Variation in the COVID-19 infection-fatality ratio by age, time and geography during the pre-vaccine era" currently under review). Third, tracking trends in the age patterns of mortality through time can help to understand local transmission dynamics and intervention impacts with greater specificity. For example, social distancing mandates like school closures, bar/restaurant closures, work-from-home orders

and restricted nursing home visitation are likely to impact the transmission risk for some age groups more than others.

The spline cascade framework has a number of limitations. First, the method for considering heterogeneous age intervals as an independent variable does not incorporate information about the population distribution within the age group. It integrates over the domain of the age group weighting each age equally. An improvement could be to incorporate population density into the optimization, and integrate over the age effect and population density simultaneously. Second, it can be difficult to choose the strength of priors. Cross-validation and simulation testing are possible solutions. Third, the underlying `MRTool` package is a linear meta-regression, not a generalized linear model, so special consideration is needed for observations at the bounds of the sampling space (e.g. observed mortality rate of zero). Fourth, because observations are used multiple times corresponding to the number of stages in the cascade, the variance of posterior uncertainty distributions may be too small. Further research is needed to ensure that posterior distributions are appropriately conservative. One potential solution is to inflate observations' measurement errors proportional to the number of stages in the cascade. Fifth, observations used for the global model are not guaranteed to be representative of the global population. Theoretically, the shape of a curve will be unbiased if other covariates in the model appropriately characterize potential confounders. In practice, accounting for confounding is difficult and unlikely to be fully achieved. Finally, as with any modeling framework, a spline cascade does not resolve underlying issues with age-disaggregated COVID-19 mortality data, such as potential under-reporting of deaths among older populations for whom cause of death attribution might be more difficult.

1.4.1 Conclusion

Bayesian spline cascades are a method for describing how non-linear patterns differ among subgroups of the data. Information about the shape of the curve is passed hierarchically by first fitting a spline-based mixed effects meta-regression using all data, then passing the estimated spline coefficients as priors to models fit on subsets of the data. This process

may be repeated for multiple levels of a hierarchy. We demonstrated how Bayesian spline cascades are a coherent solution to the challenge of data synthesis for age-specific COVID-19 deaths. Specifically, the underlying meta-regression package MRTool can take observations reported as heterogeneous age groups and infer an underlying smooth curve. The hierarchical structure of the framework, which allows information to be shared across locations, aids in adding specificity to especially wide (or missing) age groups and allows for out-of-sample predictions. Additional research is needed to further refine the method, for example, to incorporate consideration for how a population is distributed within an age group.

1.5 Bibliography

[1] IHME | COVID-19 Projections. URL: `https://covid19.healthdata.org/`.

[2] O'Driscoll, M. et al. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, pages 1–6, November 2020. Publisher: Nature Publishing Group. URL: `https://www.nature.com/articles/s41586-020-2918-0`, https://doi.org/10.1038/s41586-020-2918-0 `doi:10.1038/s41586-020-2918-0`.

[3] Excess mortality during COVID-19 in five European countries and a critique of mortality data analysis. page 13.

[4] Levin, A.T. et al. Assessing the Age Specificity of Infection Fatality Rates for COVID-19: Systematic Review, Meta-Analysis, and Public Policy Implications. *medRxiv*, page 2020.07.23.20160895, October 2020. Publisher: Cold Spring Harbor Laboratory Press. URL: `https://www.medrxiv.org/content/10.1101/2020.07.23.20160895v7`, https://doi.org/10.1101/2020.07.23.20160895 `doi:10.1101/2020.07.23.20160895`.

[5] Goldstein, J.R. and Lee, R.D. Demographic perspectives on the mortality of COVID-19 and other epidemics. *Proceedings of the National Academy of Sciences*, 117(36):22035–22041, September 2020. Publisher: National Academy of Sciences

Section: Social Sciences. URL: https://www.pnas.org/content/117/36/22035, https://doi.org/10.1073/pnas.2006392117 doi:10.1073/pnas.2006392117.

[6] MRTool, March 2021. original-date: 2020-02-11T18:21:19Z. URL: https://github.com/ihmeuw-msca/mrtool.

[7] Zheng, P. et al. Trimmed Constrained Mixed Effects Models: Formulations and Algorithms. *Journal of Computational and Graphical Statistics*, 0(0):1–13, January 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2020.1868303. https://doi.org/10.1080/10618600.2020.1868303 doi:10.1080/10618600.2020.1868303.

[8] Flaxman, A.D., Vos, T. and Murray, C.J.L. *An Integrative Metaregression Framework for Descriptive Epidemiology*. University of Washington Press, 2015. Google-Books-ID: 2676pwAACAAJ.

[9] Bigelow, J.L. and Dunson, D.B. Bayesian adaptive regression splines for hierarchical data. *Biometrics*, 63(3):724–732, September 2007. https://doi.org/10.1111/j.1541-0420.2007.00761.x doi:10.1111/j.1541-0420.2007.00761.x.

[10] Woodard, D.B., Crainiceanu, C. and Ruppert, D. Hierarchical Adaptive Regression Kernels for Regression With Functional Predictors. *Journal of Computational and Graphical Statistics*, 22(4):777–800, October 2013. URL: http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.694765, https://doi.org/10.1080/10618600.2012.694765 doi:10.1080/10618600.2012.694765.

[11] Jaeger, J. and Lambert, P. Bayesian P-spline estimation in hierarchical models specified by systems of affine differential equations. *Statistical Modelling*, 13(1):3–40, February 2013. Publisher: SAGE Publications India. https://doi.org/10.1177/1471082X12471371 doi:10.1177/1471082X12471371.

[12] Finucane, M.M. et al. Bayesian Estimation of Population-Level Trends in Measures of Health Status. *Statistical Science*, 29(1):18–25, February 2014. Publisher: Institute of Mathematical Statistics. URL: https://projecteuclid.org/journals/statistical-science/volume-29/issue-1/Bayesian-Estimation-of-Population-Level-Trends-in-Measures-of-Health/10.1214/13-STS427.full, https://doi.org/10.1214/13-STS427 doi:10.1214/13-STS427.

[13] Serra, P. and Krivobokova, T. Adaptive Empirical Bayesian Smoothing Splines. *Bayesian Analysis*, 12(1):219–238, March 2017. Publisher: International Society for Bayesian Analysis. URL: https://projecteuclid.org/journals/bayesian-analysis/volume-12/issue-1/Adaptive-Empirical-Bayesian-Smoothing-Splines/10.1214/16-BA997.full, https://doi.org/10.1214/16-BA997 doi:10.1214/16-BA997.

[14] Jonge, R.d. and Zanten, J.H.v. Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics*, 6:1984–2001, 2012. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society. URL: https://projecteuclid.org/euclid.ejs/1351603386, https://doi.org/10.1214/12-EJS735 doi:10.1214/12-EJS735.

[15] Belitser, E. and Serra, P. Adaptive Priors Based on Splines with Random Knots. *Bayesian Analysis*, 9(4):859–882, December 2014. Publisher: International Society for Bayesian Analysis. URL: https://projecteuclid.org/journals/bayesian-analysis/volume-9/issue-4/Adaptive-Priors-Based-on-Splines-with-Random-Knots/10.1214/14-BA879.full, https://doi.org/10.1214/14-BA879 doi:10.1214/14-BA879.

[16] Alexander, M. and Alkema, L. Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, 38:335–374, January 2018. Publisher: Max Planck Institute for Demographic Research. URL: https://go-gale-

com.offcampus.lib.washington.edu/ps/i.do?p=AONE&sw=w&issn=14359871&v=2.1&it=r&id=GALE%7CA568044482&sid=googleScholar&linkaccess=abs, https://doi.org/10.4054/DemRes.2018.38.15 doi:10.4054/DemRes.2018.38.15.

[17] James, G. et al. *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media, June 2013. Google-Books-ID: qcI_AAAAQBAJ.

[18] Perperoglou, A. et al. A review of spline function procedures in R. *BMC Medical Research Methodology*, 19(1):46, March 2019. https://doi.org/10.1186/s12874-019-0666-3 doi:10.1186/s12874-019-0666-3.

[19] ihmeuw-msca/mrtoolr, March 2021. original-date: 2021-03-03T20:05:41Z. URL: https://github.com/ihmeuw-msca/mrtoolr.

[20] reticulate: R Interface to Python. URL: https://rstudio.github.io/reticulate/.

[21] COVID-19 Data from the National Center for Health Statistics, April 2021. URL: https://www.cdc.gov/nchs/covid19/index.htm.

[22] Global Health Data Exchange | GHDx. URL: http://ghdx.healthdata.org/.

Chapter 2

INVESTIGATING PATTERNS OF HEALTH RISK WITH MACHINE LEARNING AND SHAP VALUES: EXAMPLES FROM THE WOMEN'S HEALTH INITIATIVE COHORT STUDY

2.1 Introduction

2.1.1 Overview

The goal of the present study is to demonstrate the utility of SHapley Additive exPlanation (SHAP) for analytic epidemiology in general and health risk prediction in particular. SHAP is a relatively new variable attribution method that quantifies the contributions of individual variables to a prediction, typically used to explain predictions from machine learning models [1]. Using data from the Women's Health Initiative cohort study (n = 93 165), we investigate how information about a subject's height, weight, hip circumference, waist circumference and age can be used to predict their mortality risk. Traditional measures of body composition, such as body-mass index and waist-to-hip ratio, are limited in their ability to characterize mortality risk due to their rigid functional form. Instead, we use a machine learning algorithm called XGBoost to summarize how anthropometric measurements are correlated with mortality risk, and we use SHAP values to investigate potential non-linear effects and interactions between the variables. Five examples are provided in the Results section.

First, we provide background on health risk prediction models and how the rise of machine learning methods has impacted the field. We then introduce variable attribution methods like SHAP, which were developed in response to the critique that most machine learning methods are perceived as "black box" algorithms. We summarize previous uses of variable attribution methods in the health field, and note a general lack of research using variable

attribution methods to gain insight into epidemiological relationships. Finally, we detail the contribution of this study, which is to demonstrate how variable attribution methods like SHAP can and should be used in the development of health risk predictions models and for analytic epidemiological research.

2.1.2 Health risk prediction and the role of machine learning

Identifying people at high risk of adverse health events is a core function of the healthcare and public health systems [2]. The concept of risk implies exposure to a hazard and a causal relationship between the hazard and the health outcome of interest [3]. Health risk prediction models aim to characterize this relationship based on individual-level data. Given the diversity of health topics and the availability of datasets, especially from electronic medical records and large cohort studies, it is perhaps not surprising that health risk prediction models are ubiquitous in the literature [4; 5; 6]. The traditional approach to health risk prediction is to use a regression model that assumes a well-defined functional relationship between predictors and the outcome, then finds coefficient values that minimize differences between the data and predictions [7]. This approach has the benefit of easily explaining why the model made a particular prediction. However, this simplicity comes with a tradeoff. Because the functional form of the model must be pre-specified, the traditional regression approach presumes that researchers already know something about how the predictors are supposed to hang together with the outcome. Is the effect of predictor x linear? Does the effect size of x depend on the value of a different predictor? Researchers must make a choice about which assumptions to encode into the functional form of a regression model, a choice that might be difficult to justify empirically [8].

While regression models are the standard in public health, they are not the only option. A class of models known as *supervised machine learning* is used for data-intensive prediction problems in several disciplines [9]. The distinction between supervised machine learning and traditional regression models is not always clear; logistic regression, for example, is sometimes considered a machine learning method. For the purpose of this article, we define machine

learning (ML) as any prediction algorithm designed to follow the information content of the data with minimal constraints on the functional relationship between the predictors and the outcome. Artificial neural networks (ANN) [10] and tree-based boosting methods [11] are commonly used algorithms that fit well within this definition. The flexibility of ML algorithms is both a strength and a limitation. If the true data generating process includes many variables with interacting and non-linear effects, ML avoids the need for a complicated model building process. Building a traditional regression model in this scenario typically amounts to an intensive guess-and-check exercise, testing and validating a wide range of prospective models. On the other hand, ML algorithms might require a large amount of data to distinguish between the information content of the data and random noise. That is, if the dataset is not sufficiently large, or for data sparse regions of the feature space, the flexibility of ML algorithms can lead to overfitting [12]. The rigidity of traditional regression models can be an asset in such data sparse conditions, when assumptions like linearity might be a more reliable basis for predictions than the data.

2.1.3 Explaining model predictions

A trained machine learning model has much to say about the information content of the data, with the capacity to recognize both general patterns and important exceptions to those patterns. For this reason, understanding why a model made a particular prediction can be prohibitively difficult, and ML in general has gained the reputation of being a set of black box methods [13]. The term "black box" refers to "anything that has mysterious or unknown internal functions or mechanisms" [14]. For domains that are especially results-oriented, like stock market prediction [15], the reason for a prediction may be less important than the fact that it is accurate. Black box models might be sufficient for these purposes. For other tasks, like consumer credit rating prediction, the reason for a prediction is crucial. The Fair Credit Reporting Act (FCRA), for example, "requires lenders to provide borrowers with relevant information about why they were given adverse or materially worse credit determinations" [16]. Given that machine learning for credit rating prediction is here to stay, these types of

regulatory requirements throw explainability of ML predictions into the spotlight. Pointing to the superior performance of a black box model in out-of-sample testing is not sufficient. As a result of this and similar applications, the development of variable attribution methods, also known as *explainability* methods, has been an active area of ML research in recent years [17; 18; 19]. Variable attribution methods (e.g. SHAP values) quantify the contributions of individual variables to a given prediction. While health risk prediction does not face the same regulatory requirements as the financial sector, model explainability is still central to the purpose of the exercise.

2.1.4 Explainability in analytic epidemiology

For certain areas of the health sciences, the reason for a prediction can even be more important than the prediction itself. Analytic epidemiology, in contrast to descriptive epidemiology (see Chapter 3), is primarily concerned with understanding the relationship between risk factors and disease outcomes [20]. The perception of ML as a black box is perhaps a reason ML is so rarely used in analytic epidemiology, because models that are truly black boxes would not be able to elucidate the individual or combined effects of risk factors. When explainability methods have been used with risk factor data, it has typically been for the task of variable selection [21; 22] or assessing the reliability of clinical models [23; 24; 25]. Using ML explainability methods to learn about risk factor effects is a nascent but growing research area, with all examples we found occurring in year 2020 or later [26; 27; 28; 29]. The recency of this development is also evidenced by lack of attention in review articles on the role of ML in public health. For example, as recently as 2018, a major review article entitled "Big Data in Public Health: Terminology, Machine Learning and Privacy" mentioned explainability methods only briefly in response to the black box critique ("recent work has partially addressed this limitation"), but did not refer to them as a way to investigate complex relationships between risk factors and health outcomes [30].

2.1.5 Goal of this study

The purpose of this article is not to claim that ML models are superior to traditional regression models for health risk prediction. That contentious debate is ongoing with both sides making valid points [31; 32; 33]. Rather, we aim to show that machine learning methods, particularly variable attribution methods like SHAP, are uniquely useful for gaining insight into patterns in the data. These methods can and should be a routine part of building health risk prediction models, to understand the information content of the data with minimal intervention by the modeler. ML explainability methods also provide an opportunity to revisit large epidemiological datasets to see what insights might have been left undiscovered by previous analyses. This perspective stands in stark contrast to the perception of ML models as black boxes, and is made possible by recent advances in ML variable attribution methods [1]. Using data from the Women's Health Initiative cohort Study [34], we explore the relationship between anthropometric measurements (e.g. height, weight, hip circumference, waist circumference, body-mass index, waist-to-hip ratio), age and mortality risk. We employ a ML algorithm called XGBoost, optimizing the Cox likelihood to account for the right-censored nature of cohort data. Finally, we use SHAP values to identify interactions and non-linear patterns among in-sample XGBoost predictions and test whether a metric based on these patterns is more predictive than body-mass index (BMI) and waist-to-hip ratio (WHR).

2.2 Methods

2.2.1 Overview

XGBoost is a tree-based gradient boosting algorithm that gained recognition for performing well in prediction competitions [35]. SHAP values quantify, for a prediction from XGBoost or another algorithm, given a set of predictor values, the additive contribution of each predictor to the prediction. The proposed method of this article is to make in-sample predictions from XGBoost or another ML algorithm, calculate a set of SHAP values for each prediction, and

derive insights from patterns in the SHAP values. This approach allows the user to directly investigate the information content of the data (as understood by the ML algorithm) as part of the model building process.

With data from the Women's Health Initiative cohort study [34], we demonstrate how this approach can uncover patterns in the data that might be difficult to observe otherwise. This is an especially important topic for right-censored cohort data in which the calculation of residuals is not straightforward. First, to demonstrate the analogy of SHAP values with the linear predictor of a regression model, Example 1 fits an XGBoost model with age and body-mass index (BMI) as predictors, and compares it to the effect sizes estimated from a traditional Cox regression model. Example 2 explores interaction effects between age and BMI using SHAP values. Example 3 demonstrates that SHAP values can be used to visualize the combined effects of multiple predictors. Example 4 quantifies the combined effects of predictor subsets, identifying which ones explain the most variation in predictions. Example 5 shows how SHAP values can highlight the limitations of pre-specified metrics like waist-to-hip ratio (WHR), and how SHAP values can be used to make new metrics of health risk. The examples progressively increase in complexity, first establishing an intuition for how SHAP values relate to more familiar concepts and later increasing the number of dimensions to demonstrate the unique capabilities that the SHAP method enables.

Users of XGBoost have control over the complexity and structure of a model by means of several hyperparameters. For simplicity, because the focus is on SHAP rather than XGBoost, we utilize the following strategy for all models: 1) to calibrate model complexity, optimize maximum tree depth and number of iterations through 5-fold cross-validation; 2) fit the a model with the resulting specification on 30 bootstrap samples of the data; 3) calculate SHAP values for the in-sample predictions from each bootstrap model; and 4) take the mean across bootstrap-specific SHAP values. While this approach does not take advantage of the full predictive power of XGBoost, which would entail optimizing several additional hyperparameters, it does provide a laptop-reproducible version sufficient for demonstrating the utility of SHAP values. All analyses were conducted using R version 4.0.2, parallelizing

XGBoost across 14 threads on a MacBook Pro with Intel Core i9 Processor (8x 2.3 GHz).

2.2.2 The XGBoost algorithm

XGBoost is a tree-based gradient boosting algorithm for making predictions from structured data [35]. For a dataset with n observations and m predictors, tree boosting models use K additive functions to predict the output: $\hat{y}_i = \psi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F}$, where $\mathcal{F} = f(x) = w_{q(x)}(q : R^m \to T, w \in \mathbb{R}^T)$ is the space of regression trees. q is a tree structure with number of leaves T and leaf weights w. Each f_k is an independent tree structure. Figure 2.1 illustrates the structure pictorially.

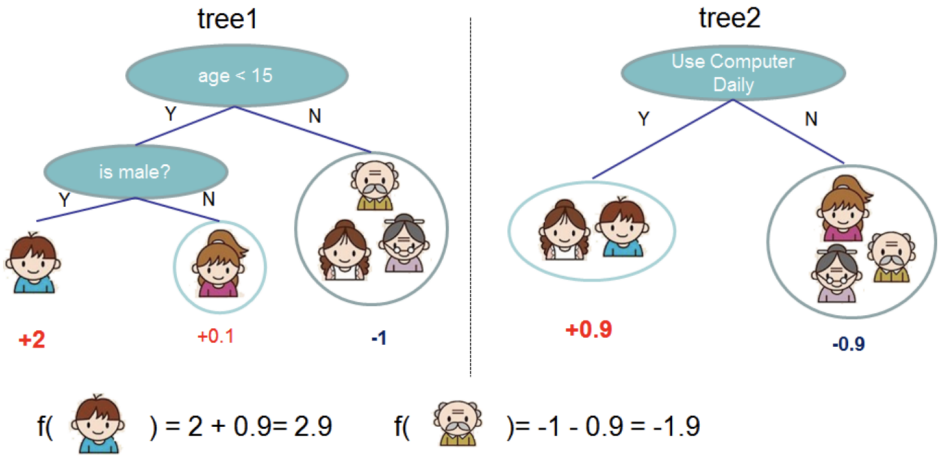

Figure 2.1: Illustration of XGBoost; 2 iterations with maximum depth 2. Predictions for observation i are made by summing the leaf weights w that contain i across each tree f_k.

To find the optimal place to bifurcate a predictor variable and grow the tree, XGBoost minimizes a loss function $\mathcal{L}(\psi)$ that includes a regularization term $\Omega(f)$:

$$\begin{aligned} \mathcal{L}(\psi) &= \sum_i l(\hat{y}_i, y_i) + \sum \Omega(f_k) \\ \Omega(f) &= \gamma T + \frac{1}{2}\lambda \|w\|^2 . \end{aligned} \tag{2.1}$$

The loss function can be a negative log likelihood as in generalized linear models, or any other differentiable convex function that quantifies the distance between the data and predictions. The XGBoost R package comes equipped with the option to handle right-censored data with Cox proportional hazards and Accelerated Failure Time likelihoods. The inputs γ (Lagrangian multiplier) and λ (L_2 regularization) are examples of user-defined hyperparameters that are often optimized through cross-validation.

The defining characteristic of boosting is to combine the contributions of many simple models (a.k.a. weak learners) to form, in the aggregate, a complex model (a.k.a. a strong learner) [26]. As a boosting algorithm, XGBoost proceeds by fitting iteration t on the residuals of earlier iterations:

$$\mathcal{L}^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}^{(t-1)} + f_t(\mathbf{x}_i)) + \sum \Omega(f_t). \tag{2.2}$$

The weak learners in this case are simple decision trees with a pre-specified maximum depth. Number of XGBoost "iterations" refers to the number of trees. Producing final predictions involves summing the predictions across decision trees [35].

XGBoost is not the only machine learning option available. In recent decades, support vector machines, Random Forest, gradient boosting machines and artificial neural networks have all been considered state-of-the-art at different points [36]. XGBoost is a natural choice due to its success in prediction competitions [35; 37]. The algorithm is approximately 10 times faster than its predecessor, gradient boosting machines [38], and a textbook about the primary alternative ML method, artificial neural networks, regards XGBoost as the de facto standard for making predictions from structured data [10]. Apart from performance, XGBoost confers some practical advantages over other ML methods. XGBoost has sustainable support from the developer community, and the implementation has built-in support for survival analysis and prediction decomposition with SHAP values [39]. These features are available in some neural network implementations, but not in a unified framework and the various options for survival analysis appear to have limited developer support [40; 41; 42]. Most importantly, XGBoost requires relatively little fine-tuning on the part of the user. In

contrast, neural networks, while flexible and powerful, require the user to specify a network's depth, structure, connectivity and other factors [10]. There is often not a straightforward way to select one among an open-ended set of options, which can be a drawback when users value reproducibility [43].

2.2.3 SHAP values for variable attribution

Model explanation methods generally fall into two categories: "global explanations that aim to explain a model's decision making process in general, and local explanations that aim to explain a single prediction" [44]. Many global variable importance methods exist, but they are typically used for the task of variable selection [45]. Local explanations like SHapley Additive exPlanations (SHAP values) are more useful for characterizing the relationship between predictors and the outcome, the purpose of this analysis. Although SHAP is a relatively new development in the field of machine learning, the method has a strong theoretical basis in game theory. The concept was originally developed in 1951 to quantify the contributions of individual players in a cooperative game. Incidentally, this problem class maps directly onto quantifying the contributions of individual predictor variables in a multivariate machine learning model.

Formally, SHAP is the average effect of removing feature j across all possible combinations of predictors (a.k.a. features) [1]:

$$\begin{aligned} g(z') &= \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \\ \phi_j &= \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{j\}) - f_x(S) \right] \end{aligned} \tag{2.3}$$

where z' is a coalition vector that indicates the presence (1) or absence (0) of each predictor in the coalition, ϕ_j is the SHAP value for feature j, ϕ_0 is the bias (i.e. intercept), N is the set of all predictor variables and S is the set of non-zero indices in z'.

Importantly, summing the predictor-specific SHAP values and the global intercept ϕ_0 recovers the predicted value. This formulation makes clear an analogy with linear regression,

which similarly describes the additive contribution of each predictor to the prediction, plus a global intercept. Consider for instance a linear regression model $E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. The contribution of predictor x_1 to the prediction $\hat{y}_i$ is $\beta_1 x_{1i}$. Or in other words, $\beta_1 x_{1i}$ can be interpreted as the SHAP value for predictor x_1 in the prediction of $\hat{y}_i$. We use this characteristic later to condition out the effect of age from the analysis. Subtracting the SHAP value for age from a prediction is analogous to subtracting out $\beta_1 age_i$ from a prediction in a linear regression.

2.2.4 Data sources and pre-processing

Data came from the Women's Health Initiative (WHI) observational cohort study [34]. The sample included 93 165 postmenopausal women ages 50 to 79, recruited between years 1993 and 1997 from 40 clinical sites in the United States. The explanatory variables used in the present study include baseline measurements of height, weight, hip circumference, and waist circumference as well as age in years. The outcome variable was a right-censored measure of time until death. Further processing of the outcome variable was not needed, because the option to optimize the Cox likelihood is a built-in feature of XGBoost. Maximum follow-up time for the WHI study was seven years. To limit the sample to participants not terminally ill at baseline, we excluded participants with follow-up time less than 365 days.

2.3 Results

2.3.1 Example 1: SHAP values are analogous to linear predictor terms in a regression model

Example 1 verifies that SHAP values from an XGBoost model capture approximately the same information as linear predictor terms from a traditional regression model. For the results shown in Figure 2.2, the XGBoost model includes only age and BMI as predictors and no constraints. The Cox regression model includes age and BMI as predictors as well, with the non-linear effect of BMI represented as a spline (p-spline with 10 degrees of freedom). The plot shows that XGBoost SHAP values (points) and a regression estimate (line) identify

approximately the same pattern of BMI's relationship with mortality, adjusting for age. This happens because SHAP values and a regression model's linear predictor share two properties. First, they are both expressed on the scale of the outcome variable. The scale of the outcome in this example is the log hazard ratio because both models optimize the Cox likelihood. Second, they are both additive. Summing a prediction's SHAP values returns the predicted value, analogous to how summing the terms of a linear predictor returns the predicted value.

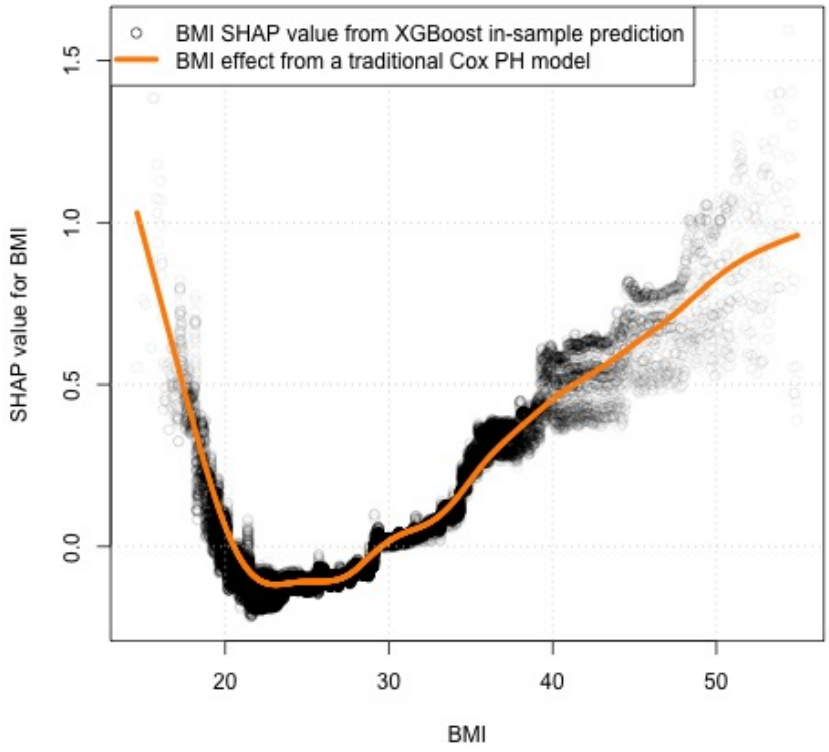

Figure 2.2: Comparison of XGBoost/SHAP and traditional Cox regression for estimating BMI effect sizes. Points are in-sample predictions' SHAP values for the BMI predictor. The orange line is the estimated BMI effect from a Cox model, on the log hazard ratio scale. Note that the orange line is estimated without reference to the points show in the plot.

2.3.2 Example 2: SHAP values are useful for identifying non-linear interactions between two predictors

Example 2 demonstrates how SHAP values can be used to identify interaction effects between variables that might be non-linear. Figure 2.3 panel (a) shows the same relationship as Figure 2.2 except now the points are colored according to age. It is apparent from the plot that the mortality risk associated with high BMI is greater for younger subjects, that is, the slope among red points is steeper than the slope among blue points. Figure 2.3 panel (b) shows the result when a traditional Cox model encodes this relationship, interacting age with each coefficient of the spline that represents BMI. The y-axis in the plot is the predicted log hazard ratio conditional on age and BMI. The Cox model recognizes the same differential effect of BMI at different ages.

2.3.3 Example 3: SHAP values help to visualize effect size patterns among many variables

While traditional regression models can easily identify interaction effects with two variables as in Example 2, the guess-and-check approach to finding interactions does not scale well with additional variables. The problem quickly becomes intractable when one cannot assume linear effects. In contrast, the machine learning approach with XGBoost and SHAP scales well provided that there is sufficient data. Example 3 considers an XGBoost model with five predictors: age, height, weight, waist circumference and hip circumference. We adjust for age by subtracting out the age SHAP value from in-sample predictions, and visualize how this predicted age-adjusted mortality risk varies by all remaining pairs of predictors in Figure 2.4.

We focus on two patterns that emerge from this way of visualizing the information. First, some plots show greater color differentiation than others, indicating that a metric based on the two variables would have better ability to distinguish between high-risk and low-risk subjects. The three pairs that include waist circumference appear to have the greatest color differentiation. We quantify this observation in Example 4. Second, the shape of some joint

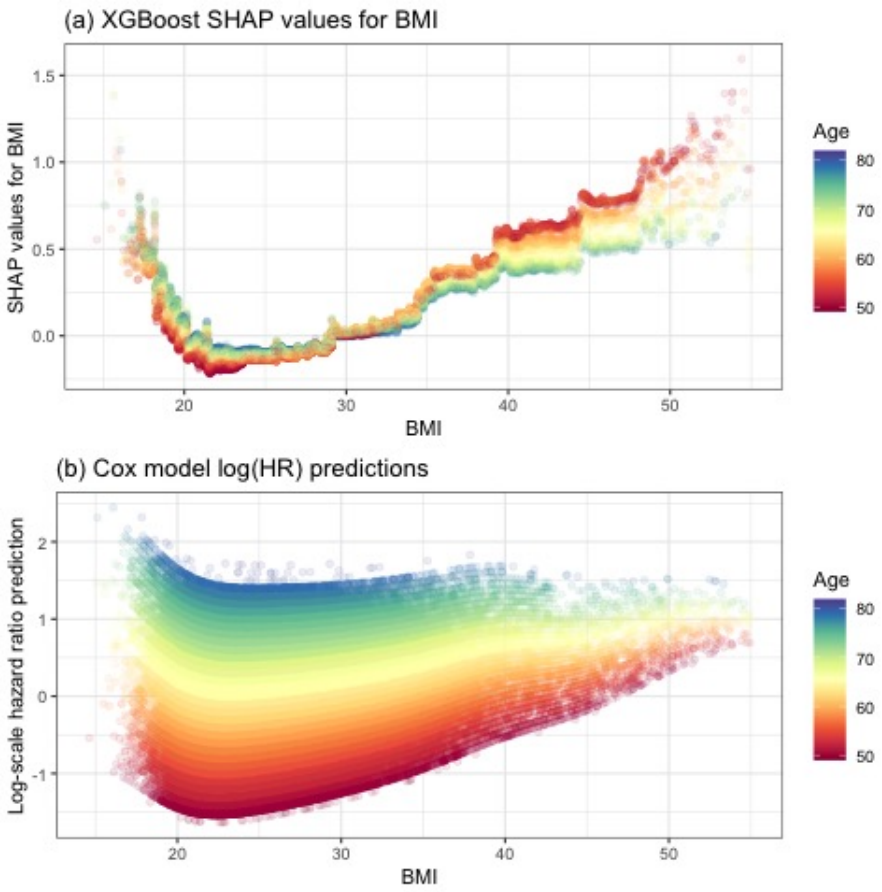

Figure 2.3: (a) The relationship between BMI and the SHAP value for BMI differs by age, indicating an interaction effect. (b) A Cox model specified to allow the non-linear effect of BMI to vary by age captures the same interaction effect.

effects is non-linear, indicating that a ratio of the two variables, which would yield a straight line, would not capture the non-linear interaction effects. This observation is relevant to body-mass index and waist-to-hip ratio as metrics of mortality risk, and we explore it further in Example 5.

2.3.4 Example 4: SHAP values can identify informative subsets of predictors

SHAP values with larger magnitude explain more heterogeneity in the dependent variable. Accordingly, summing the absolute SHAP values across multiple predictors quantifies how

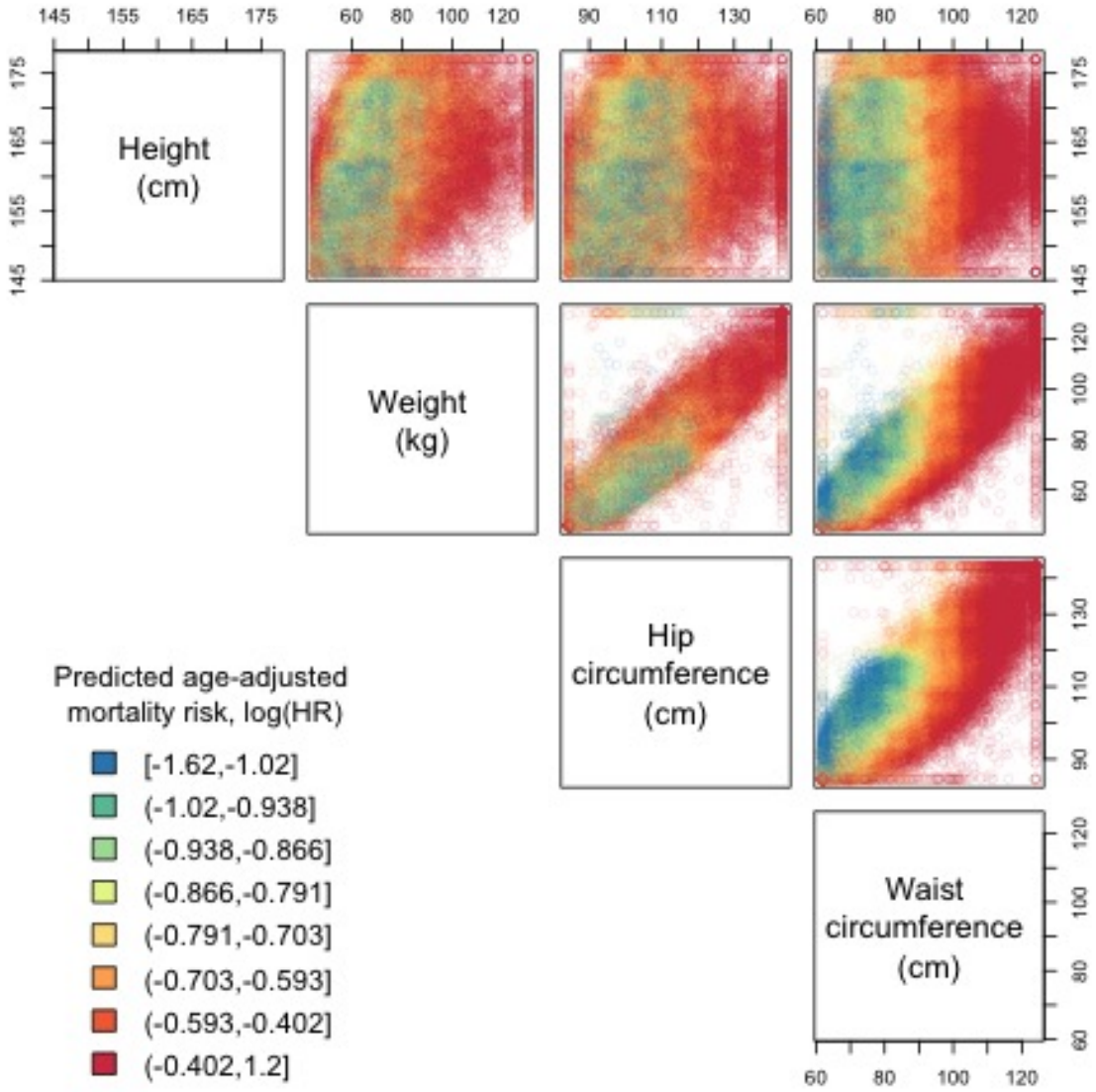

Figure 2.4: Scatterplot matrix visualizing how age-adjusted mortality risk (prediction minus age SHAP) varies across all bivariate joint distributions. One point represents one in-sample prediction.

much risk heterogeneity is explained by the combination of predictors. Absolute SHAP for

a coalition vector z' is calculated as:

$$h(z') = \sum_{j=1}^{M} |\phi_j| z'_j \tag{2.4}$$

Mean absolute SHAP, as in Table 2.1, takes the average $h(z')$ across all in-sample predictions. Percent of prediction variation explained ("% explained" in the Table 2.1) is calculated by dividing mean absolute SHAP for a set of predictors z' by the mean absolute SHAP including all predictors. The table shows that age and waist circumference alone explain 77.4% of the variation in predictions.

2.3.5 Example 5: SHAP values can be used to create new metrics of health risk

Example 5 uses SHAP values to visualize how waist-to-hip ratio fails to recognize non-linear patterns in age-adjusted mortality risk as predicted by XGBoost. The predictions shown in Figure 2.5 are from an XGBoost model with age, waist circumference and hip circumference as predictors. Colors are bins of in-sample log(HR) predictions minus the SHAP value for age (i.e. age-adjusted mortality risk). The left panel of Figure 2.5 shows that isoquants of waist-to-hip ratio do not capture non-linear patterns in age-adjusted mortality risk, as a function of waist and hip circumference. The right panel shows isoquants of a new metric that does capture the non-linear patterns. The new metric ψ is from a surrogate model predicting the sum of SHAP values for waist and hip circumference, based on independent and interaction effects between waist and hip circumference:

$$\psi_i = E[\phi_{waist_i} + \phi_{hip_i}] = \beta_0 + \beta_1 \cdot waist_i + \beta_2 \cdot hip_i + \beta_3 \cdot hip_i \cdot waist_i. \tag{2.5}$$

In 5-fold cross validation of models using ψ and $waist/hip$ as predictors, respectively, the new metric ψ yields slightly better out-of-sample discrimination and calibration than waist-to-hip ratio (average concordance: 0.6958 vs. 0.6921; average calibration coefficient: 0.99998 vs. 1.0017). The differences in performance metrics were in the expected direction

Predictors	Mean absolute SHAP	% explained
2		
age, waist	0.847	77.4
age, weight	0.698	63.9
age, height	0.641	58.6
age, hip	0.629	57.6
weight, waist	0.397	36.3
height, waist	0.340	31.1
waist, hip	0.328	30.0
height, weight	0.191	17.5
weight, hip	0.180	16.4
height, hip	0.122	11.2
3		
age, weight, waist	0.971	88.8
age, height, waist	0.913	83.6
age, waist, hip	0.902	82.5
age, height, weight	0.765	70.0
age, weight, hip	0.753	68.9
age, height, hip	0.696	63.7
height, weight, waist	0.464	42.4
weight, waist, hip	0.452	41.4
height, waist, hip	0.395	36.1
height, weight, hip	0.247	22.6
4		
age, height, weight, waist	1.038	94.9
age, weight, waist, hip	1.026	93.9
age, height, waist, hip	0.969	88.6
age, height, weight, hip	0.820	75.1
height, weight, waist, hip	0.519	47.5

Table 2.1: Mean absolute SHAP values and percent of prediction variation explained by combinations of variables

but were not statistically significant at the alpha=0.05 level. The models were:

$$\begin{aligned} h(t) &= h_0(t) + exp(\beta_1 \cdot age + pspline(metric)) \\ metric &\in \{\psi, waist/hip\}, \end{aligned} \tag{2.6}$$

where *pspline* is a p-spline with 3 degrees of freedom.

The ML-based metric did not incorporate information about weight and height, but it yielded a higher out-of-sample concordance index than a fully saturated Cox proportional hazards model with height, weight, hip and waist circumference, and age (and all of their

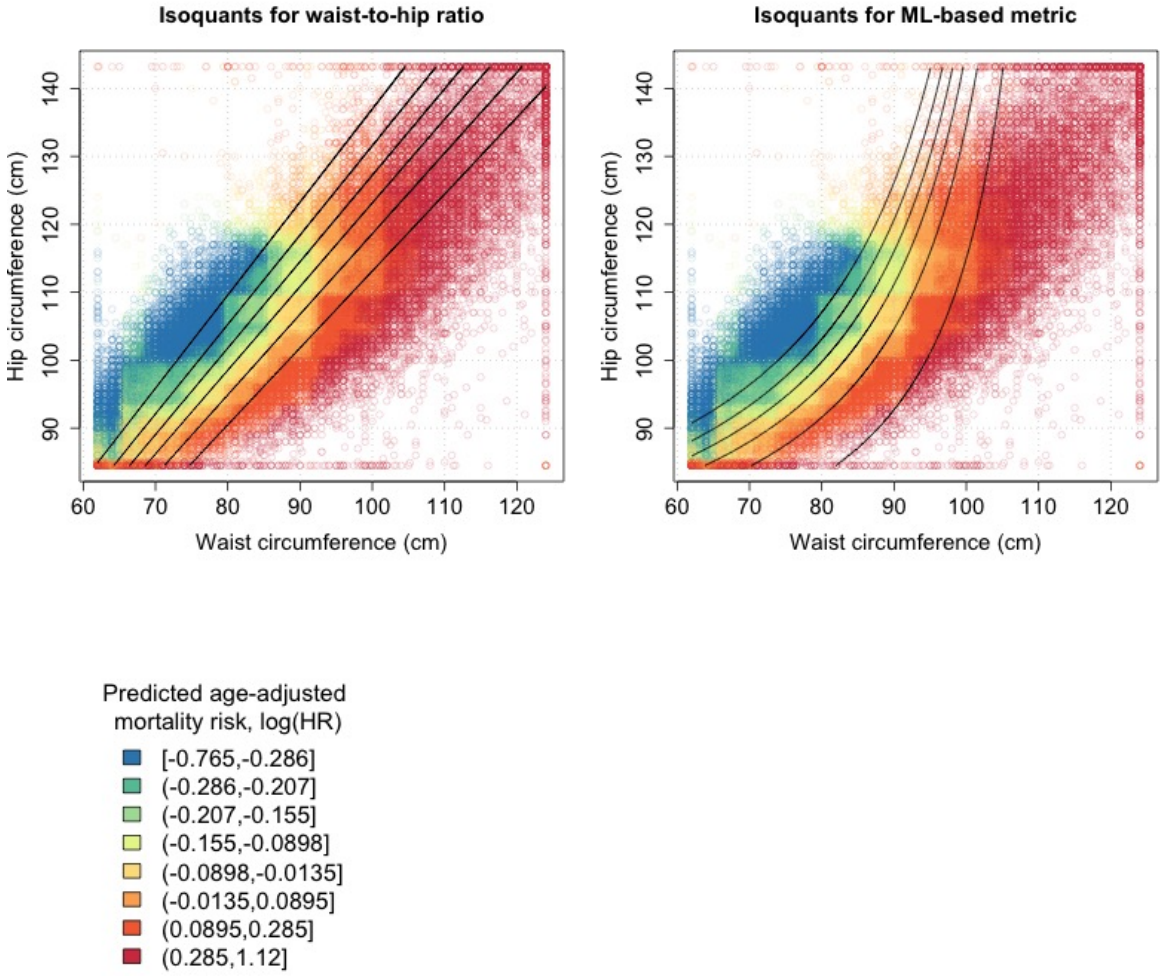

Figure 2.5: Isoquants of waist-to-hip ratio and a new ML-based metric, overlaid on the distribution of predicted age-adjusted mortality risk (in-sample prediction minus age SHAP) by waist and hip circumference

interactions) as covariates. This difference was not statistically significant, however. These results are presented in Table 2.2, along with the performance of several other Cox proportional hazards models in terms of ability to discriminate between high- and low-risk subjects. The performance of models incorporating waist and hip circumference consistently have a higher Concordance index than those incorporating measures of height and weight instead, all other model specifications being equal. This observation is consistent with the greater

color differentiation observed for waist and hip circumference in Example 3.

Model	In-sample concordance (SE)	Out-of-sample concordance (SE)
Baseline hazard only	0.5 (0)	0.5 (0)
age	0.6839 (0.003)	0.6758 (0.006)
BMI	0.5163 (0.003)	0.5287 (0.006)
WHR	0.5817 (0.003)	0.5869 (0.006)
age + BMI	0.6876 (0.003)	0.6813 (0.006)
age + WHR	0.6924 (0.003)	0.6865 (0.006)
age + pspline(BMI, 3)	0.6902 (0.003)	0.6837 (0.006)
age + pspline(WHR, 3)	0.6931 (0.003)	0.6878 (0.006)
age + pspline(BMI, 3) + pspline(WHR, 3)	0.6978 (0.003)	0.6929 (0.006)
age + weight + height + waist + hip	0.6945 (0.003)	0.6888 (0.006)
age * weight * height * waist * hip	0.6996 (0.003)	0.694 (0.006)

Table 2.2: Concordance index for Cox proportional hazards models, in-sample and out-of-sample

2.4 Discussion

2.4.1 Summary and interpretation

SHapley Additive exPlanations (SHAP values) are a useful tool in the process of developing health risk prediction models. Using data from the Women's Health Initiative cohort study, we provided examples of how SHAP values (based on XGBoost predictions) help a modeler gain qualitative and quantitative insights into the relationship between age, anthropometric measurements and mortality risk. Visualizing SHAP patterns among in-sample predictions provides a qualitative understanding about the information content of the dataset with minimal need for a priori assumptions from the modeler. Specifically, researchers can visualize effect size patterns as a function of variables of interest, adjusting for other variables (e.g. age) by subtracting out the SHAP values for those variables. Because SHAP values are expressed on the scale of the outcome variable, the absolute magnitude of SHAP values gives information about which variables are most important. When the number of predictors is

especially large, this information can help researchers quickly identify the key drivers of the outcome of interest and develop new hypotheses about the data generating process. We also demonstrated a method for using SHAP values to develop new metrics of health risk, by fitting a parametric surrogate model with SHAP values as the dependent variable. The risk metric in Example 5 outperformed waist-to-hip ratio in out-of-sample validation tests. In general, because these methods do not require a pre-specified functional form for the relationship between predictors and the outcome, they scale well with additional variables provided that there is sufficient data.

In the health sciences, data sets are often smaller than the typical use cases for machine learning (ML), as ML algorithms were primarily designed to address challenges in analyzing *big data*. Data sparsity is particularly a problem when variables are collinear. For example, two variable might encode similar information and there might not be enough data to distinguish between their individual effects. In a parametric regression framework, collinearity leads to a ballooning of the posterior uncertainty distribution and/or problems with model convergence. In contrast, XGBoost and other ML algorithms seemingly have no problem with collinearity; XGBoost does not attempt to quantify uncertainty of the predictions, and its tree-based structure means that convergence is not an issue. Has the problem simply gone away? Yes and no. The XGBoost algorithm proceeds by making binary splits in the data in a way that maximizes information gain. If two variables encode exactly the same information – information which might be useful when making predictions – the algorithm will make binary splits utilizing each of the variables with roughly equal frequency. The resulting SHAP patterns still provide important insights into the data, but the magnitude of SHAP values will be reduced corresponding to the number of collinear variables and the degree of collinearity between them. So in one sense, the XGBoost/SHAP method still provides an important way to investigate the information content of the data. In another sense, however, the fundamental problem of insufficient information has merely changed form and complicates attempts to directly interpret SHAP patterns for individual variables. It is primarily for this reason that we recommend the proposed method as a hypothesis generating exercise.

In the present analysis, for example, we discovered that waist and hip circumference are better predictors of mortality than height and weight. This finding was evident in the SHAP patterns in Example 3 and demonstrated quantitatively in the Cox proportional hazards models presented in Table 2.2; it is also consistent with the epidemiological literature, as we discuss below. Height, weight, hip circumference and waist circumference are strongly correlated and the most salient effect size patterns still came through clearly in this analysis.

2.4.2 Comparison of SHAP to other variable attribution methods

SHAP is not the only variable attribution method available, but it represents an important theoretical result that makes it a natural choice for health researchers. SHAP is the unique solution that satisfies three properties:

1. "Local accuracy: the sum of the feature attributions is equal to the output of the function;
2. Missingness: features that are missing (such that z_i'=0) are attributed no importance;
3. Consistency: changing a model so a feature has a larger impact on the model will never decrease the attribution assigned to that feature" [1].

Each of these properties is desirable when investigating health risk predictions, providing mathematical guarantees that align with intuitions from regression analysis. The property of local accuracy is particularly useful because variable contributions can be summed and subtracted without loss of generality. This property allows users to condition out the effect of a single predictor, for example, by subtracting its SHAP value from the prediction as we saw in Example 3 and Example 5.

The SHAP method unifies six previous methods for local variable attribution, including LIME, DeepLIFT, layer-wise relevance propagation and three types of classic Shapley value estimation [46]. In explaining an individual prediction, LIME fits a linear regression on model predictions within the vicinity of the feature space of the prediction [47]. This approach has

the benefit of being model-agnostic, but suffers from instability due to the limitations of linear regression (e.g. the effects of outliers) [48; 49]. Improving LIME remains an active area of research, however, and further developments may make it more reliable [50]. DeepLIFT and layer-wise relevance propagation apply only to neural networks, and the Shapley-based methods are superseded by SHAP [46].

Some common alternatives are probably better understood as visualization techniques. Partial dependence plots [51] and accumulated local effects plots [52] show the marginal effect of a feature on the outcome, averaging over all observations or observations within the vicinity of the feature space, respectively. Anchors [53], a relatively new approach from the developers of LIME, find areas in the feature space that are relatively undisturbed by changes elsewhere. These areas serve as reference points for a series of if-then statements that explain the prediction. While this is a viable, model-agnostic method that warrants further investigation, one limitation is the need to specify hyperparameter values. Finally, an open-ended option is to fit a global surrogate model on the predictions [54], fitting a simpler model with the original features as independent variables and model predictions as the dependent variable. This method improves interpretability at the cost of information content of the predictions.

2.4.3 Epidemiological considerations

Although the analysis was not intended to estimate the causal effect of body composition on mortality, the findings were consistent with several observations from the epidemiological literature. These include interaction effects between age and BMI [55], and the observation that waist and hip circumference are more informative predictors of mortality than height and weight [56]. Drawing causal conclusions about the epidemiology of obesity would require greater consideration for a number of factors. First, obesity is part of a complex network of lifestyle risk factors that are difficult to disentangle from one another [57]. Unhealthy behaviors tend to cluster together, like consumption of red and processed meats, consumption of sugar-sweetened beverages and lower physical activity [58]. A comprehensive treatment of

the topic would include these potential confounders as predictors in the model, to separate their effects from the effect of body composition. Second, observational studies are subject to reverse causation. Several health conditions (e.g. cancers, chronic renal failure and chronic obstructive pulmonary disease) decrease a person's BMI, leading to the appearance that lower BMI increases mortality risk. Researchers often exclude subjects with pre-existing conditions from the analysis for this reason, as well as remove a portion of the early follow-up period to exclude subjects with undiagnosed health problems. This analysis included a washout period of 365 days. Third, in addition to confounders and reverse causation, the presence of mediators complicates the causal relationship between body composition and mortality. Mediators are factors that lie on the causal pathway between body composition and mortality, such as blood pressure, lipids and glucose [55]. Incorporating mediators, or even proxies of mediators, as predictors can artificially decrease the effect size of interest. The validity of causal statements about body composition and mortality depends on adequately addressing these issues.

2.4.4 Limitations

This study is subject to a number of limitations. First, because the analysis considered only age and anthropometric measurements, we were not able to make substantive epidemiological comparisons between body-mass index, waist-to-hip ratio and the new ML-based metric. We made this choice to simplify the demonstration, but it came at the cost of underutilizing a rich dataset. Second, validation of a data-driven health metric is done best with data from an independent source population. BMI and WHR were at a natural disadvantage because they were not developed in reference to this particular source population as was the metric from Example 5. Third, we examined only SHAP values and did not utilize other ML explainability methods. This limited our ability to make statements about explainability methods in general, including how they compare in terms of accuracy and robustness. Fourth, we used only XGBoost and not other machine learning algorithms. It is not clear to what extent the SHAP method might have found different patterns using a different algorithm.

Fifth, as with traditional statistical models, the robustness of the proposed method depends on having sufficient data to estimate the quantities of interest. When variables are roughly collinear, for example, there might not be enough information to distinguish their individual effects. For XGBoost and SHAP, this situation manifests as a sharing of the effect size across multiple variables, which complicates the interpretation of SHAP patterns for individual variables. Finally, we did not take full advantage of newer methodological developments in the SHAP method, like the calculation of SHAP interaction values. This could have strengthened the demonstration by showing functionality that can be applied to additional use cases. The common thread in these limitations is the relatively small scope of the analysis. Based on the promising results, however, we believe further research expanding the scope to include additional covariates, datasets, machine learning algorithms, explainability methods and SHAP functionality is warranted.

2.4.5 Conclusion

In conclusion, Shapley Addition exPlanations (SHAP) values and similar methods for explaining machine learning predictions are underutilized in the field of analytic epidemiology. Contrary to the popular perception of ML algorithms as "black box" methods, we showed that ML algorithms like XGBoost can be used to gain insight into the information content of the data in a way that is simple, scalable and minimizes the need for a priori modeling assumptions. Specifically, using data from the Women's Health Initiative cohort study, we demonstrated how SHAP values can be used to identify salient patterns for predicting mortality, and how that information can be translated into a measure of health risk. Because SHAP values aid in learning about epidemiological relationships, they can and should be a routine part of developing health risk prediction models.

2.5 Bibliography

[1] Lundberg, S.M., Erion, G.G. and Lee, S.I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]*, February 2018. arXiv: 1802.03888.

URL: http://arxiv.org/abs/1802.03888.

[2] Moons, K.G.M. et al. Prognosis and prognostic research: what, why, and how? *BMJ*, 338:b375, February 2009. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting. URL: https://www.bmj.com/content/338/bmj.b375, https://doi.org/10.1136/bmj.b375 doi:10.1136/bmj.b375.

[3] Council, N.R. et al. *Science and Decisions: Advancing Risk Assessment*. National Academies Press, March 2009. Google-Books-ID: FS3txQ5C0WcC.

[4] Damen, J.A.A.G. et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*, 353:i2416, May 2016. Publisher: British Medical Journal Publishing Group Section: Research. URL: https://www.bmj.com/content/353/bmj.i2416, https://doi.org/10.1136/bmj.i2416 doi:10.1136/bmj.i2416.

[5] Kourou, K. et al. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015. https://doi.org/10.1016/j.csbj.2014.11.005 doi:10.1016/j.csbj.2014.11.005.

[6] Goldstein, B.A. et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198–208, January 2017. https://doi.org/10.1093/jamia/ocw042 doi:10.1093/jamia/ocw042.

[7] Dunn, P.K. and Smyth, G.K. *Generalized Linear Models With Examples in R*. Springer, November 2018. Google-Books-ID: tBh5DwAAQBAJ.

[8] Harrell, F.E. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer series in statistics. Springer, Cham Heidelberg New York, second edition edition, 2015. OCLC: 922304565.

[9] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3):160, March 2021. https://doi.org/10.1007/s42979-021-00592-x doi:10.1007/s42979-021-00592-x.

[10] Chollet, F. and Allaire, J.J. *Deep Learning with R*. Manning Publications Company, 2018. Google-Books-ID: xnIRtAEACAAJ.

[11] James, G. et al. *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media, June 2013. Google-Books-ID: qcI_AAAAQBAJ.

[12] Hawkins, D.M. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, January 2004. Publisher: American Chemical Society. https://doi.org/10.1021/ci0342472 doi:10.1021/ci0342472.

[13] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Computer science;Criminology;Science, technology and society;Statistics Subject_term_id: computer-science;criminology;science-technology-and-society;statistics. URL: https://www.nature.com/articles/s42256-019-0048-x, https://doi.org/10.1038/s42256-019-0048-x doi:10.1038/s42256-019-0048-x.

[14] Definition of BLACK BOX. URL: https://www.merriam-webster.com/dictionary/black+box.

[15] Shah, D., Isah, H. and Zulkernine, F. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studies*, 7(2):26, June 2019. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. URL: https://www.mdpi.com/2227-7072/7/2/26, https://doi.org/10.3390/ijfs7020026 doi:10.3390/ijfs7020026.

[16] Chou, A. WHAT'S IN THE BLACK BOX? BALANCING FINANCIAL INCLUSION AND PRIVACY IN DIGITAL CONSUMER LENDING. *Duke Law Journal*, 69(5):1183–1218, February 2020. Publisher: Duke University, School of Law. URL: https://go-gale-com.offcampus.lib.washington.edu/ps/i.do?p=AONE&sw=w&issn=00127086&v=2.1&it=r&id=GALE%7CA616047476&sid=googleScholar&linkaccess=abs.

[17] Adadi, A. and Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. Conference Name: IEEE Access. https://doi.org/10.1109/ACCESS.2018.2870052 doi:10.1109/ACCESS.2018.2870052.

[18] Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267, June 2017. https://doi.org/10.1016/j.artint.2018.07.007 doi:10.1016/j.artint.2018.07.007.

[19] Guidotti, R. et al. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51, February 2018. https://doi.org/10.1145/3236009 doi:10.1145/3236009.

[20] Principles of Epidemiology | Lesson 1 - Section 7, February 2019. URL: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section7.html.

[21] Madakkatel, I. et al. Can we use machine learning to discover risk factors? Testing the proof of principle using data on >11,000 predictors and mortality in the UK Biobank. *medRxiv*, page 2021.05.07.21256791, May 2021. Publisher: Cold Spring Harbor Laboratory Press. URL: https://www.medrxiv.org/content/10.1101/2021.05.07.21256791v1, https://doi.org/10.1101/2021.05.07.21256791 doi:10.1101/2021.05.07.21256791.

[22] Hu, L. et al. Tree-Based Machine Learning to Identify and Understand Major Determinants for Stroke at the Neighborhood Level. *Journal of the Ameri-*

can Heart Association: Cardiovascular and Cerebrovascular Disease, 9(22):e016745, November 2020. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7763737/, https://doi.org/10.1161/JAHA.120.016745 doi:10.1161/JAHA.120.016745.

[23] Amann, J. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):NA–NA, November 2020. Publisher: BioMed Central Ltd. URL: https://go-gale-com.offcampus.lib.washington.edu/ps/i.do?p=AONE&sw=w&issn=14726947&v=2.1&it=r&id=GALE%7CA650641218&sid=googleScholar&linkaccess=abs, https://doi.org/10.1186/s12911-020-01332-6 doi:10.1186/s12911-020-01332-6.

[24] Cutillo, C.M. et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *npj Digital Medicine*, 3(1):1–5, March 2020. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Diagnosis;Disease prevention;Medical imaging;Public health;Therapeutics Subject_term_id: diagnosis;disease-prevention;medical-imaging;public-health;therapeutics. URL: https://www.nature.com/articles/s41746-020-0254-2, https://doi.org/10.1038/s41746-020-0254-2 doi:10.1038/s41746-020-0254-2.

[25] Thorsen-Meyer, H.C. et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2(4):e179–e191, April 2020. Publisher: Elsevier. URL: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30018-2/abstract, https://doi.org/10.1016/S2589-7500(20)30018-2 doi:10.1016/S2589-7500(20)30018-2.

[26] Li, R. et al. Machine Learning–Based Interpretation and Visualization of Nonlinear

Interactions in Prostate Cancer Survival. *JCO Clinical Cancer Informatics*, (4):637–646, November 2020. Publisher: Wolters Kluwer. URL: https://ascopubs.org/doi/10.1200/CCI.20.00002, https://doi.org/10.1200/CCI.20.00002 doi:10.1200/CCI.20.00002.

[27] Moncada-Torres, A. et al. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, 11:6968, March 2021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7998037/, https://doi.org/10.1038/s41598-021-86327-7 doi:10.1038/s41598-021-86327-7.

[28] Peng, J. et al. An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients. *Journal of Medical Systems*, 45(5):61, April 2021. https://doi.org/10.1007/s10916-021-01736-5 doi:10.1007/s10916-021-01736-5.

[29] Jansen, T. et al. Machine Learning Explainability in Breast Cancer Survival. *Digital Personalized Health and Medicine*, pages 307–311, 2020. Publisher: IOS Press. URL: https://ebooks.iospress.nl/doi/10.3233/SHTI200172, https://doi.org/10.3233/SHTI200172 doi:10.3233/SHTI200172.

[30] Mooney, S.J. and Pejaver, V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health*, 39(1):95–112, 2018. _eprint: https://doi.org/10.1146/annurev-publhealth-040617-014208. https://doi.org/10.1146/annurev-publhealth-040617-014208 doi:10.1146/annurev-publhealth-040617-014208.

[31] Kelly, C.J. et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, October 2019. https://doi.org/10.1186/s12916-019-1426-2 doi:10.1186/s12916-019-1426-2.

[32] Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction mod-

els. *Journal of Clinical Epidemiology*, 110:12–22, June 2019. Publisher: Elsevier. URL: https://www.jclinepi.com/article/S0895-4356(18)31081-3/abstract, https://doi.org/10.1016/j.jclinepi.2019.02.004 doi:10.1016/j.jclinepi.2019.02.004.

[33] Maddox, T.M., Rumsfeld, J.S. and Payne, P.R.O. Questions for Artificial Intelligence in Health Care. *JAMA*, 321(1):31–32, January 2019. https://doi.org/10.1001/jama.2018.18932 doi:10.1001/jama.2018.18932.

[34] Group, T.W.H.I.S. Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1):61–109, February 1998. https://doi.org/10.1016/s0197-2456(97)00078-0 doi:10.1016/s0197-2456(97)00078-0.

[35] Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[36] Zhang, L. et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 22(11):1680–1685, November 2017. URL: http://www.sciencedirect.com/science/article/pii/S1359644616304366, https://doi.org/10.1016/j.drudis.2017.08.010 doi:10.1016/j.drudis.2017.08.010.

[37] Nielsen, D. Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition? Master's thesis, NTNU, 2016.

[38] Xgboost presentation. URL: https://cran.r-project.org/web/packages/xgboost/vignettes/xgboostPresentation.html.

[39] GitHub - dmlc/xgboost: Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more. Runs on sin-

gle machine, Hadoop, Spark, Flink and DataFlow. URL: https://github.com/dmlc/xgboost.

[40] Liu, P. liupei101/TFDeepSurv, November 2019. original-date: 2018-11-01T03:14:49Z. URL: https://github.com/liupei101/TFDeepSurv.

[41] Ching, T. cox-nnet: Extension of neural networks for Cox Regression. URL: https://github.com/traversc/cox-nnet.

[42] Wang, P., Li, Y. and Reddy, C.K. Machine Learning for Survival Analysis: A Survey. *arXiv:1708.04649 [cs, stat]*, August 2017. arXiv: 1708.04649. URL: http://arxiv.org/abs/1708.04649.

[43] McDermott, M.B.A. et al. Reproducibility in Machine Learning for Health. *arXiv:1907.01463 [cs, stat]*, July 2019. arXiv: 1907.01463. URL: http://arxiv.org/abs/1907.01463.

[44] van der Linden, I., Haned, H. and Kanoulas, E. Global Aggregations of Local Explanations for Black Box models. *arXiv:1907.03039 [cs]*, July 2019. arXiv: 1907.03039. URL: http://arxiv.org/abs/1907.03039.

[45] Wei, P., Lu, Z. and Song, J. Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, 142:399–432, October 2015. URL: http://www.sciencedirect.com/science/article/pii/S0951832015001672, https://doi.org/10.1016/j.ress.2015.05.018 doi:10.1016/j.ress.2015.05.018.

[46] Lundberg, S.M. and Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Guyon, I. et al, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[47] Ribeiro, M.T., Singh, S. and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM. event-place: San Francisco, California, USA. URL: http://doi.acm.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778 doi:10.1145/2939672.2939778.

[48] Molnar, C. *5.7 Local Surrogate (LIME) | Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/lime.html#fn38.

[49] Alvarez-Melis, D. and Jaakkola, T.S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[50] Lee, E. et al. Developing the sensitivity of LIME for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100610. International Society for Optics and Photonics, May 2019. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/1100610/Developing-the-sensitivity-of-LIME-for-better-machine-learning-explanation/10.1117/12.2520149.short, https://doi.org/10.1117/12.2520149 doi:10.1117/12.2520149.

[51] Molnar, C. *5.1 Partial Dependence Plot (PDP) | Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/pdp.html.

[52] Molnar, C. *5.3 Accumulated Local Effects (ALE) Plot | Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/ale.html.

[53] Ribeiro, M.T., Singh, S. and Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*, 2018.

[54] Molnar, C. *5.6 Global Surrogate | Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/global.html.

[55] Hu, F. *Obesity Epidemiology*. Oxford University Press, March 2008. Google-Books-ID: rv42OY9QLkMC.

[56] Sahakyan, K.R. et al. Normal-Weight Central Obesity: Implications for Total and Cardiovascular Mortality. *Annals of Internal Medicine*, 163(11):827–835, December 2015. https://doi.org/10.7326/M14-2525 doi:10.7326/M14-2525.

[57] van Dam, R.M. et al. Combined impact of lifestyle factors on mortality: prospective cohort study in US women. *The BMJ*, 337, September 2008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2658866/, https://doi.org/10.1136/bmj.a1440 doi:10.1136/bmj.a1440.

[58] Schulze, M.B. et al. Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women. *JAMA*, 292(8):927–934, August 2004. https://doi.org/10.1001/jama.292.8.927 doi:10.1001/jama.292.8.927.

Chapter 3

COMPARING METHODS FOR ESTIMATING UNDER-5 MORTALITY IN 193 COUNTRIES: A SIMULATION STUDY

3.1 Introduction

3.1.1 Overview

The quantity of interest for the present validation study is the probability of dying before age five, also known as ${}_5q_0$ or under-5 mortality. Using country-level under-5 mortality data from the Global Burden of Disease (GBD) study [1], including 44 936 observations from 193 countries spanning the period since 1950, we created several simulated datasets in which the underlying true values are known. We then tested several modeling strategies on their ability to discover truth based on noisy data with various patterns of missingness. The models include a hierarchical Bayesian spline model (introduced in Chapter 1), a machine learning algorithm called XGBoost (introduced in Chapter 2), and linear and non-linear mixed effects models [2; 3]. We also tested using each of these models as the first stage of a modeling framework called Spatio-temporal Gaussian Process Regression (ST-GPR), commonly utilized as part of the GBD study [4]. Finally, we tested two approaches for combining results from the various models to improve overall performance, a process known as *ensemble modeling* [5].

We first describe the process of model validation, including goals, methods and special considerations for descriptive epidemiological models. We emphasize the importance of domain-specific knowledge in the design of validation studies, for example, choosing realistic scenarios in which to test the models and choosing performance metrics that correspond to the intended use of the model. We then summarize previous research on validating descriptive epidemiological models and describe the contribution of this study.

3.1.2 Model validation

Model validation can be defined as "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses" [6]. Model validation is a key step in the development of descriptive epidemiological models, which attempt to characterize a population in terms of some measure of health loss (e.g. disease prevalence, mortality rates). As the roots of the word epidemiology suggest – the study of what is upon ("epi") the people ("demos") [7] – these sorts of descriptive models are a core component of the field of epidemiology and its role in society [8]. Results from descriptive epidemiological models are a critical input for decisions about health policy, healthcare resource allocation [9], targeted interventions [10], global health philanthropy [11] and the prioritization of public health research [12]. Decision-makers rely on these results to give an accurate description of health problems in a population, and tracking how epidemiological measures change through time is often an important step toward understanding the drivers of population health patterns [13]. Validating descriptive epidemiological models is thus not a mechanical exercise. Because the purpose of these models is action-oriented, the process requires thinking critically about what consumers of the information value in a model and how best to communicate the reliability of results.

3.1.3 Simulation and data holdouts

Model validation takes many forms corresponding to domain-specific threats to model validity. Threats often include certain patterns of data sparsity or outliers unique to a particular data-generating process. The field of econometrics has an extensive body of research testing the performance of estimators, or functions that map a sample space to a set of sample estimates, in real-world conditions [14]. In contrast to mathematical proofs describing how estimators behave with infinitely large sample sizes and regular probability distributions, econometricians typically use Monte Carlo simulation to observe how estimators perform with small sample sizes and distributional irregularities. The process of Monte Carlo simula-

tion entails "1) modeling the data-generating process, 2) generating several sets of artificial data, 3) employing these data and an estimator to create several estimates, and 4) using these estimates to gauge the sampling distribution properties of that estimator for the particular data-generating process under study" [14]. Domain-specific knowledge is needed for modeling the data-generating process because reasonable values for the parameters must be specified a priori.

Another model validation technique, one common in the field of machine learning, is to use cross-validation and other forms of data holdouts. In contrast to Monte Carlo simulation where the true underlying distribution of the data is known, it is possible to use portions of the real training dataset to approximate the true underlying distribution [15]. The model might be fit on 80 percent of the data, for example, and predictions on the other 20 percent are compared to the observed values according to some distance metric like mean absolute error. Here as well, domain-specific knowledge is useful in determining holdout patterns that match the patterns of missingness expected in the applied setting.

3.1.4 Uncertainty quantification

The distinction between data holdouts and Monte Carlo simulation is important for uncertainty quantification. A confidence interval aims to capture uncertainty in the estimation of a parameter's true value, whereas a prediction interval aims to capture the expected variation in hypothetical new observations. In statistical terminology, the distinction is variance of the expectation of Y (confidence interval) versus variance of Y itself (prediction interval), conditional on covariates [16]. Misunderstandings of this distinction are common in the literature. For example, one commentator offered the following critique of a set of global health estimates: "The differences between survey results and global health estimates are not trivial. [...] In six countries, the survey statistics were outside the uncertainty ranges of the estimates (two for UN and four for IHME)" [17]. This observation confuses the difference between a confidence interval and a prediction interval, as it is not expected that 95 percent of observations fall within the 95 percent uncertainty interval. The same misunder-

standing occurs when researchers attempt to use cross-validation to validate their approach to confidence interval quantification. If the task is to estimate the true value of a parameter, validation of an uncertainty interval must be in reference to the underlying true value, not other observations in the dataset. For this reason, validation studies for uncertainty quantification typically use a simulation approach in which the underlying truth is known.

Uncertainty quantification is an important aspect of descriptive epidemiological models because the true data generating mechanism is often complex and subject to many ambiguities. According to Mathers and colleagues, "Uncertainty in estimated disease burden may arise from the following sources: incomplete information, for example, when estimates for a population are based on observations from a sample; potential biases in information, for instance, issues concerning the representativeness for a whole population of estimates from a study of a subgroup or the validity of a survey instrument in addressing the quantity of interest; heterogeneity or from disagreements among information sources, as when several studies give different estimates for the same quantity of interest; model uncertainty, for example, the variables or functional form specified in a regression model; the data generation process itself; for instance, investigators may only infer risks from event counts in a population, which means that they can never know the risks themselves with certainty" [18]. In the field of global health, where data can vary substantially in quantity and quality across locations, capturing differences in the strength of evidence is critical to making informed decisions. If a model fails to incorporate the relevant sources of uncertainty into the final estimates, the resulting narrow uncertainty can lead to overconfidence in a model's results and misallocation of resources.

3.1.5 Choosing performance metrics

The selection of performance metrics is a statement about which characteristics of a model are most valuable. Referring back to the definition of model validation as "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses" [6], the concept of "intended uses" implies a value

judgment on the part of model users. They attach utility to aspects of how the model behaves within the context of their research. To the degree that users can articulate what they value in a model and operationalize it as a metric, they can determine quantitatively how acceptable a model is for its intended use. The preceding section is an example of such a judgment. If the purpose of the model is to discover the true value of a parameter, the relevant performance metric for uncertainty is how often the confidence interval encompasses the true underlying value. If the purpose of the model is to predict the range within which future observations are likely to occur, coverage of the prediction interval is the relevant performance metric.

Some model characteristics, such as high accuracy and precision, are valuable simply by virtue of the definition of modeling, which is to create an accurate representation of reality based on imperfect information. We hope for the representation to be close to the truth on average (high accuracy), and we prefer for individual predictions to be closer to the truth rather than farther away (high precision). These concepts are sometimes called bias and variance, or systematic error and random error. Even in this fundamental case, however, value judgements are required based on intended use of the model. We provide two examples. First, there are multiple ways to operationalize precision of predictions with respect to the true value, for example, mean absolute error (MAE) and root mean square error (RMSE) [19; 20]. RMSE increases quadratically as errors increase, whereas MAE increases linearly. If users hope to avoid especially large errors in the applied setting, even if it means a greater frequency of smaller errors, RMSE is perhaps a better choice as a performance metric. Second, models with high accuracy do not always have high precision, and vice versa. In fact, the opposite tends to be true, that there is a fundamental tradeoff between accuracy and precision, known as the "bias-variance" tradeoff in the machine learning literature [15]. Ridge regression, for example, is known not to converge on the true value with more data but individual predictions tend not to be egregiously wrong [21]. This characteristic may be desirable for researchers social sciences, for example, who may have datasets with lower sample sizes and a high degree of unexplained variation in the quantity of interest.

Accuracy and precision may be fundamental desirable qualities in a model, but their relative importance depends on the model's intended use.

Accuracy, precision and coverage of the uncertainty interval are not the only characteristics that matter. Researchers commonly also value low computational cost, statistical efficiency, and asymptotic properties like consistency and normality [14].

3.1.6 Non-sampling error and other threats to model validity

Attempts to discover the true values of population parameters, like disease prevalence, are complicated by a number of factors. Collecting information from each person in a population is often prohibitively expensive and time-consuming, so survey studies must select a subset of people to represent the wider population. Assuming that the subset is a faithful microcosm of the wider population, the difference between prevalence in the subset and true prevalence in the population is known as the *sampling error* [22]. This quantity is unknowable without conducting a full census, but likely values of the sampling error can be described as a probability distribution. In practice, many additional and sometimes unmeasurable factors come into play. For example, diagnostic tools for assessing a person's disease status are imperfect, and attempts to obtain a representative sample of the population are thwarted by people's refusal or inability to participate in the survey. These are examples of *non-sampling error* [22]. Additional contributors to the uncertainty of a result include uncertainty in the covariates used to aid in prediction [23] (i.e. "errors-in-variables"), the selection of one among many valid estimation procedures [24] (i.e. model uncertainty), and other contributors to total error [25].

When non-sampling error is not accounted for in an individual study, it biases the result upward or downward according to the magnitude of the error. In meta-analysis, which considers multiple studies, non-sampling error is an additional source of variability which may or may not bias the mean estimate. Meta-analysis is the process of synthesizing results from multiple studies to obtain an overall estimate for the effect of some factor on an outcome of interest. The distinction between fixed-effects meta-analysis and random-effects meta-

analysis is relevant for the discussion of non-sampling error. The defining assumption of fixed-effects meta-analysis is that there exists a single true effect size that all component studies are attempting to estimate. The assumption implies that all studies would converge on the same estimated value as the sample sizes of the studies approach infinity. In contrast, random-effects meta-analysis relaxes this assumption and allows for variation in the true effects [26]. As sample sizes hypothetically go toward infinity, it is not assumed that studies would converge on the same estimate. An additional parameter describes the distribution of between-study effects, specifically the variance of a Normal distribution constrained to have a mean value of zero. Random-effects meta-analysis is the standard approach in descriptive and analytic epidemiology because, in practice, there are a wide variety of reasons studies may differ from each other. We return to this topic later when introducing the candidate models for the present study, some of which are mixed effects models that estimate a term for the non-sampling variation, or the degree of between-study heterogeneity. Because sources of non-sampling error differ across research disciplines, domain-specific knowledge is needed to anticipate the ways in which anomalies (variation not due to low sample size or the effects of measured variables) can present themselves in a dataset. Insofar as these anomalies represent a threat to model validity, a validation study needs to demonstrate the robustness of a model to those threats. For descriptive epidemiology, there are three characteristics of a dataset that potentially affect the validity of modeled results: data sparsity, data comparability and data quality. These are related to but different than the sources of uncertainty outlined above in Section 3.1.4.

- Data sparsity refers to a paucity or lack of information about the quantity of interest for a population, where a population is defined by location, time and potentially other characteristics like age and sex. Two common examples of data sparsity are missing data in recent years and missing data in resource-poor areas. Finding up-to-date data is a perennial problem, and resource-poor areas tend to have the worst health problems. Taken together, this means that information is often the least plentiful for the time

periods and locations most relevant for public health action.

- Data comparability refers to the degree of consistency in the definition of the quantity of interest across data sources. Low data comparability can result from differential biases between sampling strategies (e.g. phone-based or door-to-door surveys), using disparate case definitions, and using diagnostic tests with different sensitivity and specificity. A common remedy is for researchers to convert estimates from non-standard measurement methods into the equivalent gold standard, a process known as "cross-walking," but there might not be sufficient information to make such a conversion.

- Data quality refers to the degree of fidelity in the process of measuring the quantity of interest and representing the information in the dataset. For meta-analyses in descriptive epidemiology, data quality issues can stem from human error at many points along the information pipeline. These include errors during primary data collection, statistical analysis to obtain a study's result, reporting results inaccurately or without enough context in the literature, and during the process of extracting results from the literature. Data quality is also affected by measurement instruments with low sensitivity, low specificity or low inter-rater reliability.

3.1.7 Previous validation studies for descriptive epidemiology

Among modeling frameworks for descriptive epidemiology, we found two that describe validation methods in detail. The Cause of Death Ensemble model (CODEm) [27] aims to estimate cause-specific mortality rates using data from vital registration systems and verbal autopsies. For validation, Foreman and colleagues [27] fit a model on 70 percent of the data and tested it on the other 30 percent using Root Mean Square Error (RMSE), a test of trend and coverage of the prediction interval as validation metrics. Their approach has strengths and limitations. First, a strength is the rigorousness of their data holdout strategy for assessing the threat of data sparsity. They create two test sets that reflect the missing-

ness by age and year in the input data. The first test set is used to calibrate a smoothing parameter, and final out-of-sample predictive validity is quantified using the second test set. Second, the authors extensively describe threats to model validity from data quality and comparability. They note issues related to "mapping across various revisions of the ICD, variation in garbage coding across countries and time, misclassification due to poor diagnostic capacity, comparability of alternative verbal autopsy methods, and completeness of cause of death registration, and large nonsampling variance." While they cite outlier identification as important for these reasons, and describe potential ways to deal with outliers, they do not incorporate this domain-specific knowledge into the model validation process itself. Third, the validation metrics did not include quantification of the direction of bias, which could have been captured with mean error, for example. Finally, the choice of prediction interval coverage as a validation metric does not capture what users need to know about uncertainty. The purpose of the model is to describe true cause-specific mortality rates in a population within a margin of error. This is the role of a confidence interval, the validity of which cannot be assessed by comparing predictions to other observations in the dataset.

The second model is DisMod-MR [28], which aims to estimate several health-related parameters simultaneously (e.g. incidence, prevalence, remission, etc.), taking into consideration the mathematical relationships between them. Using epilepsy and schizophrenia datasets, Flaxman and colleagues employed cross-validation to compare different ways of specifying rate models, including parameterizing them as Poisson, negative-binomial and other distributions. Using 100 random subsets of the source data, they fit models on 75 percent of the data and compared it to the other 25 percent. First, a limitation is that the authors selected source datasets that are relatively homogeneous in age and geography in order to minimize the effect of random noise on the validation study. However, this makes the results of the study less generalizable to other models that might vary substantially by age and geography, which the authors note. Second, as with CODEm, they used a metric that describes how often an interval encompassed observations, rather than truth. This approach does not capture uncertainty in the same sense as a confidence interval. The authors

considered Monte Carlo simulation to address this limitation but expressed concern that the choice of probability distribution for the data-generating mechanism would favor some models over others.

3.1.8 Goals and characteristics of this study

The goal of the present validation study is to assess alternative strategies for modeling under-5 mortality rates for all countries during the period since 1950. After defining a data generating mechanism that reflects the essential characteristics of the under-5 mortality dataset used in the GBD study, we generate five random realizations for each of four missingness scenarios that are commonly encountered in the field of global descriptive epidemiology. We do this to improve the generalizability of results to other models. Using only information about a population's monetary wealth, level of maternal education, HIV mortality rates and year of observation, each model predicts child mortality rates in every country for each year between 1950 and 2019. Model performance is quantified for location-years with data in the original dataset according to Root Mean Square Error (RMSE), mean error and coverage of the 95 percent confidence interval. All performance metrics assess predictions against the underlying true value. Two of the candidate models are new, proposed frameworks: Bayesian spline cascades were introduced in Chapter 1, and the XGBoost algorithm was introduced in Chapter 2. We elaborate on the specific implementations of these approaches in the Methods section. In addition to assessing the models in isolation, we use the models as first-stage predictions in Spatio-temporal Gaussian Process Regression (ST-GPR). ST-GPR is one of three modeling frameworks at the Institute for Health Metrics and Evaluation (IHME) used for making global estimates that are comprehensive across locations, age groups, sexes and years since 1990 (and sometimes earlier). The others are CODEm and DisMod, the models with validation studies highlighted in Section 3.1.7. ST-GPR is described in greater detail in the Methods section. Finally, we assess combinations of model predictions in an effort to improve overall model performance, a process known as *ensemble modeling*.

This analysis builds on the strengths of previous validation studies in a number of ways.

First, this study uses Monte Carlo simulation rather the data holdouts or cross-validation. This allows models to be assessed against ground truth rather than other observations in a dataset, an essential characteristic for quantification of uncertainty intervals. Second, we create simulated datasets that reflect the types of data sparsity expected in the applied setting for under-5 mortality modeling and other descriptive epidemiological models. The sparsity types include no missing data, data missing in recent years, data missing for a period in the middle of the time series and completely holding out data from a subset of locations. Third, we adjusted under-5 mortality observations that were derived from measurement methods other than established reference methods (typically vital registration systems and complete birth histories). This is a process that can be implemented as a standard part of data preparation for modeling in the applied setting. The non-sampling variation remaining after adjustment reflects the type of unexplained variation expected in the applied setting. Fourth, this study uses a larger and more geographically comprehensive set of observations than previous validation studies. Fifth, we assess predictions independently and as part of ST-GPR, which effectively quantifies the added benefit conferred by ST-GPR for the first time in simulation. Sixth, we present two novel modeling strategies for under-5 mortality: the Bayesian spline cascade allows for fine-grained control over the functional form of the model, while the machine learning algorithm (XGBoost) allows for a high degree of flexibility. In combination with linear and non-linear mixed effects models, this diversity of modeling strategies allows for a deeper discussion about the role of constraints and flexibility in descriptive epidemiological modeling. Seventh, we assess the benefit conferred by ensemble modeling with very diverse models. Whereas previous approaches made ensembles of the same model with different covariate sets or different types of mixed effect regressions, the candidate models in this study represent completely different modeling paradigms.

3.2 Methods

3.2.1 Overview

This simulation study has five key steps. First, with the set of survey series, vital registration systems and other data sources that comprise the Global Burden of Disease child mortality dataset (hereafter called only "survey series"), we define a data generating mechanism consistent with the observed data. This process includes adjusting non-standard child mortality measurements to their equivalent gold standard values, estimating a mean function among all locations and years with data, and characterizing within- and between-series variability in a location. Second, we create five random realizations of the data generating mechanism for each of four missingness scenarios: 1) missing years after 2000, 2) missing years 1990 to 1999, 3) missing 20 percent of locations, and 4) no missing data. Third, on each of the 20 resulting datasets, we fit four canditate models on log-transformed under-5 mortality rates using covariates for lag-distributed income, mean years of education among women age 15 to 49, and HIV death rate in the under-5 population. The candidate models were a linear mixed effects model (Model A), a non-linear mixed effects model designed to roughly approximate the current GBD approach (Model B), a Bayesian spline cascade where the time effect is represented as a B-spline (Model C), and XGBoost fit on first-differences within survey series (Model D). Fourth, we used each of the resulting 80 sets of predictions as the first stage of an ST-GPR model to obtain 80 distinct sets of estimates with uncertainty for all GBD locations during years 1950 to 2019. We also ensembled the four models' predictions. Fifth, we assessed all predictions against the underlying true mean function in terms of root mean squared error, mean error and coverage of the uncertainty interval.

3.2.2 Data preparation

The simulations in this study are based on the dataset used in the Global Burden of Disease (GBD) study for estimating under-5 mortality [1]. The dependent variable is mortality rate among children aged under 5 years, or ${}_5m_0$, denoted here simply as m. After excluding

information from subnational locations, the dataset included 44 936 national-level observations from 1,977 survey series in 193 countries. Observations are indexed by country (c), year (t) and survey series (s). To account for that fact that some methods of measuring child mortality are more accurate than others, we adjusted non-reference survey series to a country-specific reference as defined in the GBD model. The GBD model chooses complete vital registration systems as the reference where available, and otherwise chooses the best available survey or combination of surveys by location [1]. We conducted the adjustment by finding all instances where a non-reference (a) and reference (r) observation overlap on location and year, taking the difference in log-scale under-5 mortality rates for this pair, and modeling this difference using a mixed effects regression with random intercepts. The random intercepts for the first adjustment were country effects nested within region nested within non-reference measurement method type k (Eq. 3.1):

$$\begin{aligned} \delta_{1k,c} &= (\log(m_a) - \log(m_r))_{c,k} = \beta_0 + u_k + u_{k:region} + u_{k:region:c} \\ \log(m^{'}_{c,t,s,i}) &= \log(m_{c,t,s,i}) - \hat{\delta}_{1k,c}, \end{aligned} \tag{3.1}$$

where the $'$ asterisk indicates that an adjustment has occurred.

Subsequently, to reduce the potential for compositional bias within a location, we conducted a second adjustment intercept-shifting individual survey series s to the reference, modeling random intercepts as country nested within survey series (Eq. 3.2):

$$\begin{aligned} \delta_{2s,c} &= (\log(m_a) - \log(m_r))_{c,s} = \beta_0 + u_s + u_{s:c} \\ \log(m^{''}_{c,t,s,i}) &= \log(m^{'}_{c,t,s,i}) - \hat{\delta}_{2s,c}. \end{aligned} \tag{3.2}$$

All random intercept terms u are assumed to follow a Gaussian distribution with mean 0 and variance estimated from the data. By conducting these adjustments as intercept shifts, the method assumes that the degree of bias for a given non-reference survey is consistent through time.

3.2.3 Defining the data generating mechanism and creating simulated datasets

To make simulated datasets, we needed three pieces of information for each location c: the mean function describing log-scale child mortality rates over time $\tau(t)$, the variance of between-series heterogeneity γ, and the variance of within-series heterogeneity σ^2 (Eq. 3.3).

$$\begin{aligned} \log(m_{c,t,s,i}) &\sim \tau(t)_c + u_s + \epsilon_i \\ u_s &\sim N(0, \gamma_c) \\ \epsilon_i &\sim N(0, \sigma_c^2) \end{aligned} \tag{3.3}$$

By location, we first fit models to the data that has been adjusted for survey method and survey series. The nonlinear mean function for describing change in child mortality over time $\tau(t)_c$ comes from a thin-plate spline model using the *gamm* function in the R *mgcv* package [29]. The function optimizes the degree of smoothness through generalized crossvalidation, and estimates random intercepts by survey series. The location-specific variance of between-series heterogeneity γ_c comes from the estimated variance of the random intercepts for a given location. The location-specific variance of within-series heterogeneity σ_c^2 is calculated as the average of squared survey-specific residuals, where residuals are the differences from predictions incorporating the survey-specific random intercepts (Eq. 3.4):

$$\sigma_c = \Sigma_{i=1}^{n} \frac{\log(m_{c,t,s,i}^{''}) - (\log(\hat{\tau}(t)_c) + \hat{u}_s)}{n} \tag{3.4}$$

We then created five random realizations of the data-generating process for each location, with data reflecting the exact coverage of each survey series in the original data. For example, if a real survey was conducted annually in Angola between 2002 and 2019, each random realization would also contain a simulated survey in Angola with coverage for the years 2002 through 2019. For each of these five simulated datasets, we created four holdout patterns typical of missingness encountered in global health datasets: missing recent years (missing years after 2000), missing years in the middle of a time series (missing 1990 through 1999), missing entire locations (missing 20 percent of locations), and the complete dataset

(no missingness). This process resulted in 20 simulated datasets that reflect the essential characteristics of the original data.

3.2.4 Candidate models

Next, we fit four candidate models on each of the 20 datasets. The information available to the models was year t of the under-5 mortality observation, location c of the observation (including country, region and super-region identifiers), and lag-distributed income per capita (denoted LDI), mean years of education among women aged 15 to 49 years (denoted *education*), and HIV death rate in the under-5 population (denoted HIV) [1] all indexed by location and time of the observation. Note that these covariates were not involved in the process of creating the true data generating mechanism for each location.

Model A was a linear mixed effects model (Eq. 3.5) fit using the *lmer* function from the R *lme*4 package [2].

$$\begin{aligned} \log(m_{c,t,s,i}) &= \mathbf{x}\beta + u_c + u_s + \epsilon_i \\ u_c &\sim N(0, \gamma_C) \\ u_s &\sim N(0, \gamma_S) \\ \epsilon_i &\sim N(0, \sigma^2), \end{aligned} \tag{3.5}$$

where m_{ct} is the under-5 mortality rate, $\mathbf{x}\beta$ is the linear predictor, u_c is the location-specific random intercept, u_s is the survey-specific random intercept, and ϵ_i represents measurement error.

Model B was a non-linear mixed effects model [3] designed to roughly mirror the current GBD approach (Eq. 3.6). The primary difference is that data derived from alternative methods of measuring under-5 mortality have already been adjusted to the reference in this simulation study, whereas the GBD model adjusts them as part of the non-linear mixed effects model. This difference was necessary to maintain comparability of the candidate models, namely that the other models (other than Model B) did not need to conduct an equivalent in-model adjustment in order to be comparable to Model B.

$$
\begin{aligned}
&m_{c,t,s,i} = \exp[\beta_0 + u_{0cs} + (\beta_1 + u_{1c})\text{LDI}_{ct} + (\beta_2 + u_{2c})\text{education}_{ct}] + (\beta_3 + u_{3c})\text{HIV}_{ct} + \epsilon_{c,t,s,i} \\
&u_{0cs} \sim N(0, \gamma_{u_{0cs}}) \\
&u_{1c} \sim N(0, \gamma_{u_{1c}}) \\
&u_{2c} \sim N(0, \gamma_{u_{2c}}) \\
&u_{3c} \sim N(0, \gamma_{u_{3c}}) \\
&\epsilon_{c,t,s,i} \sim N(0, \sigma^2),
\end{aligned}
\tag{3.6}
$$

where each u term is a Gaussian distributed random effect (slope or intercept) with variance denoted by γ, $\epsilon_{c,t,s,i}$ is the residual, and the covariates were defined earlier in this section. Predictions incorporate the location-specific random effects but not the source-specific random effects.

Model C was a Bayesian spline cascade with time represented as a spline and the LDI, *education* and HIV covariates represented as linear fixed effects. The cascade included four stages: global, super-region, region and location. Every model included random intercepts by location, except for models that contained data for only one location. In this case, the variance of random intercepts was constrained to be zero so as not to overidentify the quantity estimated by the fixed effect intercept. The spline on time was a quadratic B-spline, had three internal knots spaced according to data density and had linear tails. In each stage, we re-estimated the spline coefficients with Gaussian priors coming from the parent model in the hierarchy. For the non-spline covariates, we estimated coefficient values in the global model and then set them as constant for the subsequent stages of the cascade.

Following the mathematical notation developed in Chapter 1, the global model fits a spline on time and estimates coefficients for the fixed effects as in a usual meta-regression:

$$
\begin{aligned}
y_{ij}^1 &= (x_{ij}^1)^T \beta_j^1 + (z_{ij}^1)^T \zeta_j^1 + \epsilon_{ij}^1 \\
\epsilon_{ij}^1 &\sim N(0, (\sigma_j^1)^2),
\end{aligned}
\tag{3.7}
$$

where superscript 1 ($k = 1$) indicates the first stage of the cascade, j indicates a group in

level k (which for the first stage is only one possible group), x is the design matrix for the spline, β is the vector of coefficients corresponding to spline bases, z is the design matrix for the three covariates (LDI, *education* and HIV), ζ is the vector of coefficients corresponding to the covariates, and ϵ is the residual.

The subsequent stages of the cascade re-estimate β using the parent model's estimates coefficients as a Gaussian prior, with the variance of the prior V scaled by user-defined λ which has been set to 1 for all stages of the cascade. The ζ values remain constant as those estimated in the first stage:

$$\begin{aligned} \hat{\beta}_j^k &= \arg\min_{\beta\in\Omega} \sum_{i=1}^{N_j} g(\beta; y_{ij}^k, \Sigma_j^k) + p(\beta; \hat{\beta}_{j_\uparrow}^{k-1}, \lambda^k V_{j_\uparrow}^{k-1}) \\ y_{ij}^k &= (x_{ij}^k)^T \beta_j^k + (z_{ij}^k)^T \hat{\zeta}_j^1 + \epsilon_{ij}^k, \end{aligned} \tag{3.8}$$

where the first formula is the general definition of the negative log likelihood in a cascade model (same as Eq. 1.1 in Chapter 1), the second formula indicates that ζ has been set to the estimated values from stage 1, and $j_\uparrow$ refers to the parent model specific to model j.

Model D was the XGBoost algorithm fit on annualized differences within survey series:

$$\Delta_{c,s,t} = \frac{\log(m_{c,s,t_2}) - \log(m_{c,s,t_1})}{t_2 - t_1}, \tag{3.9}$$

where m_{l,s,t_2} and m_{l,s,t_1} refer to two observations from the same survey series that are adjacent in time. We divide the difference in rates by the number of years occurring between the two surveys to calculate the average annual difference in under-5 mortality during the period. The resulting annualized difference $\Delta_{l,s,t}$ is indexed in time to the midpoint of the time interval t.

For each of 30 bootstrap samples of the data, we fit an XGBoost model optimizing maximum tree depth and number of iterations through 5-fold cross-validation. The effects of LDI and *education* were constrained to be monotonically decreasing, and the effect of HIV was constrained to be monotonically increasing. Covariates for super-region, region and location were passed to the model using target encoding [30]. The country covariate was independent while region, super-region, LDI, *education* and HIV were allowed to freely

interact with each other. This allows region-specific effects of HIV on under-5 mortality, for example, but not country-specific effects. The predicted annualized difference for location c in year t is the mean of bootstrap predictions:

$$\hat{\Delta}_{c,t} = \sum_{b=1}^{30} \psi(\mathbf{y}_b, \mathbf{x}_b)/30 \tag{3.10}$$

where ψ is the XGBoost algorithm described above. We obtained a final time series estimate by predicting annual differences for all years in a location, taking the cumulative sum, and intercept-shifting the resulting relative curve such that squared differences with observed data in the most detailed geography available are minimized:

$$\begin{aligned} \log(\hat{m}_{c,t}) &= \pi_c + \log(\hat{m}^r_{c,t}) \\ \log(\hat{m}^r_{c,t}) &= \sum_{t'<t} \hat{\Delta}_{c,t'} \\ \pi_c &= \sum_{i=1}^{n} (\log(m_{g,t,i}) - \log(\hat{m}^r_{c,t}))^2/n \\ g &\in \{country, region, superregion\}, \end{aligned} \tag{3.11}$$

where $\log(\hat{m}^r_{c,t})$ is a country's relative time series at year t and π_c is the intercept that brings a country's relative time series back to the level of the observed data in geography g.

Model E was an ensemble of Models A, B, C and D, taking the average prediction for each combination of location and year for years 1950 through 2019. Model F was an ensemble of predictions that result from using Models A, B, C and D as the first stage predictions in Spatiotemporal Gaussian Process Regression (ST-GPR). With 250 posterior draws from each component model, Model F obtained the point estimate by taking the unweighted mean of the pooled draws and obtained the uncertainty interval by taking the 2.5% and 97.5% quantiles of the pooled draws.

3.2.5 Spatiotemporal Gaussian Process Regression (ST-GPR)

ST-GPR is a modeling framework designed to estimate a full time series for each country and several subnational regions, typically stratified by sex and age group [4]. In the first

stage of ST-GPR, the default method is to employ a mixed-effects linear regression to make predictions for GBD demographics defined by age, sex, location and year where applicable. Country-level and study-level covariates are incorporated in this stage and are not needed in subsequent stages. Alternative models may be used for the first stage. Residuals of this model are then smoothed by age, year and location in the "space-time" (ST) stage. User-defined hyperparameters control the degree of residual smoothing in each of three dimensions: λ for time, ω for age and ζ for sharing information across levels of the geographical hierarchy. The λ parameter is defined as: $w_{c,a,s,t} = \frac{1}{e^{\lambda|t-t_0|}}$, where w is the weight, t is a point in time, and t_0 is another point in time. The ζ parameter is defined as $w_{c,a,s,t} = \zeta^{|c-c_0|}$, where ζ^0 is for country data, ζ^1 is for regional data not from the country being estimated, ζ^2 is for data from other regions in the same super-region, and ζ^2 is for global data from other super regions. For purposes of the simulation, we optimized the values of λ and ζ through cross-validation using adaptive grid search to settle on final values. ω has no influence in this particular model because only one age group is modeled. The hyperparameter values used in this analysis were $\lambda = 0.2$, $\zeta = 0.02$, and $\omega = 1$.

Formal definitions of these hyperparameters can be found in the methods appendix to the GBD 2019 risk factors paper [4]. Finally, Gaussian process regression (GPR), a non-parametric and Bayesian form of regression, further smooths the estimates and produces uncertainty intervals. The GPR step takes as a prior the mean prediction $m_{c,a,s}(t)$ with Matern covariance:

$$g_{c,a,s}(t) \sim GP(m_{c,a,s}(t), Cov(m_{c,a,s}(t))) \tag{3.12}$$

. Matern covariance is defined as:

$$M(t,t') = \sigma^2 \frac{2^{1-v}}{\Gamma(v)} (\frac{d(t,t')\sqrt{2v}}{l})^v K_v (\frac{d(t,t')\sqrt{2v}}{l}) \tag{3.13}$$

where σ^2 is the marginal variance, estimated as normalized absolute deviation between the first-stage prediction and the smooth residuals; v is degree of differentiability, a smoothness parameter set as a constant; and l is length scale, a user-defined parameter roughly meaning the distance at which points become uncorrelated.

The mean function is defined as:

$$m_{c,a,s}(t) = X_{c,a,s}\beta + h(r_{c,a,s,t}), \tag{3.14}$$

where $X_{c,a,s}\beta$ is the linear predictor from a mixed effect linear regression (which may be replaced with another model) and $h(r_{c,a,s,t})$ is the smoothed residual for a population defined by country, age group, sex and time [4]. In the current version of ST-GPR, variance of the input data ("data variance") for each observation is considered in the GPR step but not the first stage or residual smoothing steps. In the simulation, data variance for each observation is set to be the country-specific within-survey variance used to generate the simulated data.

3.2.6 Assessing model performance

Finally, we assessed all predictions against the underlying true mean function in terms of root mean squared error (RMSE), mean error and coverage of the confidence interval. RMSE assesses how closely the predictions align with the data generating mechanism that produced the simulated datasets:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{|S|}\sum_{c,t\in S} e_{c,t}^2} \\ e_{c,t} &= E(y|\mathbf{x}_{c,t}) - \tau(t)_c, \end{aligned} \tag{3.15}$$

where $E(y|\mathbf{x}_{c,t})$ is the model prediction conditional on covariates and $\tau(t)_c$ is the underlying true log-scale under-5 mortality rate in country c at time t.

Mean error maintains the sign of the difference between predictions and truth, which enables a determination for whether a model systematically overestimates or underestimates the true value:

$$ME = \frac{1}{|S|}\sum_{c,t\in S} e_{c,t}. \tag{3.16}$$

Coverage is, for a given set of predictions based on simulated data in which the true underlying value is known, the proportion of true values that lie within the uncertainty interval. For a 95 percent uncertainty interval, the optimal value is 0.95.

3.3 Results

3.3.1 Overview

We report metrics quantifying how well the candidate models performed in the task of discovering the true under-5 mortality rate, using noisy simulated data in various scenarios where data are missing for select years and locations. Based on patterns that appeared in the quantitative results, we investigate five observations further by looking at examples from various countries and more detailed performance metrics. The detailed performance metrics include stratifying results by whether or not a location had only a single data source (always a vital registration system), and results for an ensemble approach that takes draws from the posterior GPR distributions for each component model.

Performance metrics are stratified by whether a unique location-year combination was in-sample or out-of-sample for a given missingness scenario (Figure 3.1). For simplicity, we

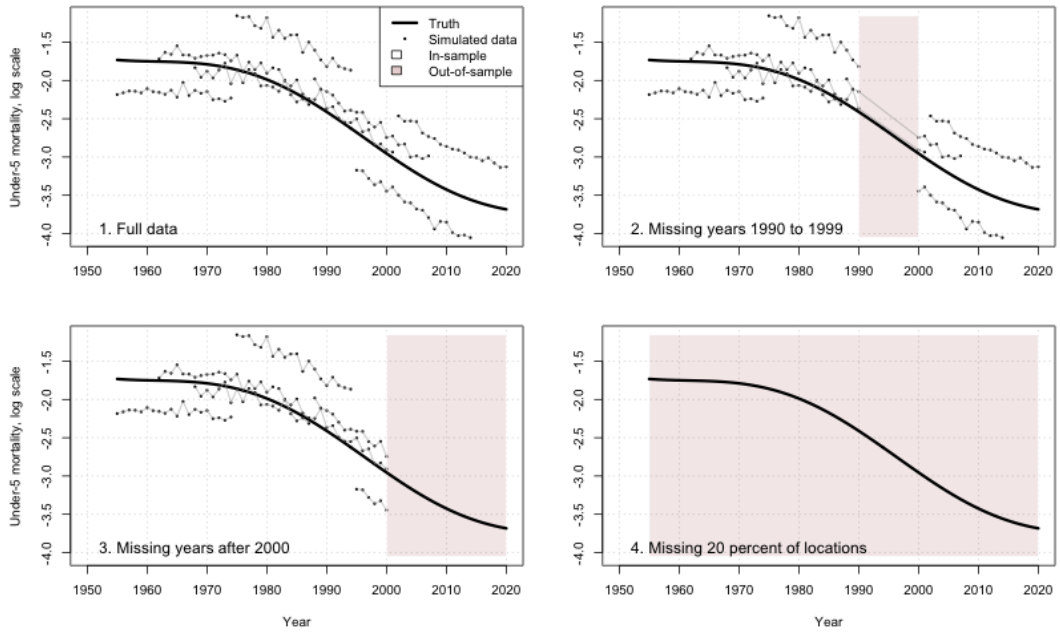

Figure 3.1: Examples of the four missingness scenarios for simulated data in a given location; performance metrics are reported separately for in-sample and out-of-sample populations

call a unique location-year combination a *population*. In-sample populations are those that contained data for the model that defined the data generating mechanism (and therefore received a value that represents truth) and also had simulated data in the missingness scenario. Out-of-sample populations are those that contained data for the model that defined the data generating mechanism but did not have simulated data in the missingness scenario. The models were exposed only to the simulated data in a given missingness scenario (e.g. the black points), and were assessed on how well they predicted the underlying true curve (black line in Figure 3.1).

In addition to stratification by in-sample and out-of-sample populations, we also stratify results by whether the prediction came directly from a particular candidate model ("first stage"), after space-time smoothing ("ST"), or after the GPR step ("GPR"). Note that uncertainty is reported only for GPR outputs. Performance metrics for out-of-sample populations are not applicable to the "Full data" scenario and are not reported.

3.3.2 Quantitative results, out-of-sample metrics

Table 3.1 shows performance metrics among out-of-sample prediction populations under three missingness scenarios. Considering first stage predictions, the ensemble model (Model E) had the lowest RMSE in all three scenarios. Among the non-ensemble models, the spline cascade (Model C) had the lowest RMSE in all scenarios and had much lower bias than the next best alternative when extrapolating into the future (mean error = 0.0001 compared to -0.0205). XGBoost (Model D) had the highest RMSE in two scenarios and second-highest in the time extrapolation scenario. The poor performance of XGBoost in the scenario with missing locations is somewhat expected because the model was primarily designed to estimate a relative time trend, and the intercept shift is based on the most detailed geography with data.

Considering the effect of ST smoothing, all models had lower RMSE than their respective original predictions used in the first stage. The ensemble model had the lowest RMSE in two scenarios and the second-lowest in the other scenario. The spline cascade model had

Out-of-sample performance metrics

	First stage		ST		GPR		
Model	RMSE	Mean error	RMSE	Mean error	RMSE	Mean error	Coverage
Missing 20 percent of locations							
Model A: Linear ME	0.5447	0.0482	0.5101	0.0899	0.5101	0.0899	0.422
Model B: Nonlinear ME	0.5960	-0.0393	0.5893	-0.0270	0.5893	-0.0270	0.246
Model C: Spline cascade	0.5338	0.1443	0.5273	0.1825	0.5273	0.1825	0.427
Model D: XGBoost	1.3756	-1.0774	1.3589	-1.0431	1.3589	-1.0431	0.107
Model E: Ensemble of point predictions	0.3963	-0.0314	0.3811	-0.0070	0.3811	-0.0070	0.378
Model F: Ensemble of draws	NA	NA	NA	NA	0.4660	-0.1994	0.911
Missing years 1990 to 1999							
Model A: Linear ME	0.2078	0.0437	0.1347	0.0534	0.1321	0.0653	0.710
Model B: Nonlinear ME	0.2171	0.0147	0.1350	0.0486	0.1357	0.0654	0.639
Model C: Spline cascade	0.2009	0.0602	0.1453	0.0700	0.1378	0.0730	0.595
Model D: XGBoost	0.2219	0.0675	0.1425	0.0540	0.1386	0.0649	0.662
Model E: Ensemble of point predictions	0.1735	0.0538	0.1310	0.0552	0.1300	0.0659	0.610
Model F: Ensemble of draws	NA	NA	NA	NA	0.1319	0.0672	0.704
Missing years 2000 and later							
Model A: Linear ME	0.4152	0.0380	0.3520	0.0365	0.3344	0.0431	0.458
Model B: Nonlinear ME	0.5309	-0.0274	0.4247	0.0096	0.4106	0.0241	0.368
Model C: Spline cascade	0.3405	0.0001	0.2965	0.0259	0.2897	0.0331	0.490
Model D: XGBoost	0.4682	-0.0205	0.4463	-0.0082	0.4450	-0.0004	0.389
Model E: Ensemble of point predictions	0.3292	0.0299	0.3129	0.0402	0.3086	0.0476	0.397
Model F: Ensemble of draws	NA	NA	NA	NA	0.3137	0.0249	0.656

Table 3.1: Performance metrics calculated among out-of-sample prediction populations, average of five random realizations

the lowest RMSE in the scenario requiring extrapolation forward in time. Regarding mean error, many of the models in scenarios missing periods of time were biased upward, the only exception being XGBoost having a slight downward bias in the "Missing years 1990 to 1999" scenario.

GPR further decreased RMSE in all cases except one, but the amount of decrease was substantially less compared to the effect of ST smoothing. Note that ST and GPR are the same by definition in the "Missing 20 percent of locations" scenario. Coverage of the uncertainty interval was lower than 0.95 in all scenarios. The model with the highest coverage was the linear mixed effects model in the scenario with missing data for years 1990 to 1999

In-sample performance metrics

	First stage		ST		GPR		
Model	RMSE	Mean error	RMSE	Mean error	RMSE	Mean error	Coverage
Full data							
Model A: Linear ME	0.2645	-0.0046	0.1234	0.0340	0.1220	0.0542	0.647
Model B: Nonlinear ME	0.2458	-0.0002	0.1210	0.0340	0.1223	0.0526	0.600
Model C: Spline cascade	0.2215	0.0097	0.1208	0.0401	0.1201	0.0538	0.612
Model D: XGBoost	0.2027	0.0022	0.1162	0.0386	0.1204	0.0536	0.592
Model E: Ensemble of point predictions	0.1776	0.0125	0.1152	0.0379	0.1197	0.0531	0.591
Model F: Ensemble of draws	NA	NA	NA	NA	0.1203	0.0535	0.628
Missing 20 percent of locations							
Model A: Linear ME	0.2579	-0.0033	0.1239	0.0313	0.1234	0.0518	0.656
Model B: Nonlinear ME	0.2404	0.0012	0.1206	0.0308	0.1229	0.0497	0.596
Model C: Spline cascade	0.2270	0.0118	0.1217	0.0359	0.1219	0.0515	0.636
Model D: XGBoost	0.1925	0.0031	0.1154	0.0358	0.1203	0.0510	0.593
Model E: Ensemble of point predictions	0.1764	0.0134	0.1158	0.0346	0.1206	0.0505	0.597
Model F: Ensemble of draws	NA	NA	NA	NA	0.1213	0.0510	0.637
Missing years 1990 to 1999							
Model A: Linear ME	0.2738	-0.0093	0.1243	0.0350	0.1237	0.0530	0.672
Model B: Nonlinear ME	0.2712	-0.0032	0.1232	0.0354	0.1246	0.0518	0.632
Model C: Spline cascade	0.2127	0.0066	0.1222	0.0418	0.1215	0.0523	0.629
Model D: XGBoost	0.2273	0.0030	0.1200	0.0404	0.1209	0.0531	0.639
Model E: Ensemble of point predictions	0.1865	0.0110	0.1181	0.0392	0.1214	0.0522	0.625
Model F: Ensemble of draws	NA	NA	NA	NA	0.1217	0.0526	0.657
Missing years 2000 and later							
Model A: Linear ME	0.2063	-0.0065	0.0966	0.0269	0.0981	0.0428	0.655
Model B: Nonlinear ME	0.1385	0.0093	0.0891	0.0252	0.0969	0.0409	0.619
Model C: Spline cascade	0.1890	0.0187	0.0960	0.0357	0.0968	0.0435	0.620
Model D: XGBoost	0.1489	-0.0043	0.0890	0.0266	0.0935	0.0399	0.563
Model E: Ensemble of point predictions	0.1305	0.0108	0.0881	0.0295	0.0939	0.0414	0.587
Model F: Ensemble of draws	NA	NA	NA	NA	0.0952	0.0418	0.637

Table 3.2: Performance metrics calculated among in-sample populations, average of five random realizations

(coverage = 0.71). The ensemble (Model F) increased coverage in all cases except one. Coverage in general was best (closest to 0.95) when predicting for years missing in the middle of the time series, next best when extrapolating forward in time, and worst when predicting for entirely missing locations.

3.3.3 Quantitative results, in-sample metrics

Table 3.2 shows performance metrics among in-sample prediction populations under four missingness scenarios including the full data scenario. Considering the first stage predictions, RMSE was lowest for the ensemble model (Model E) and highest for the linear mixed effects model (Model A) in all scenarios. Mean error was most negative for the linear mixed effects model but was generally small in all cases. The non-linear mixed effects model (Model B) and XGBoost were the least biased on average in terms of absolute mean error. Among non-ensemble models, XGBoost generally had the lowest RMSE, with the exceptions that the spline cascade model had slightly lower RMSE in the "Missing years 1990 to 1999" scenario and the non-linear mixed effects model had slightly lower RMSE in the "Missing years 2000 and later" scenario. XGBoost and the non-linear mixed effects model had the least bias on average in terms of absolute mean error.

Considering the effect of ST smoothing, all models had lower RMSE than the original predictions used as each of their respective first stages. Compared to first stage predictions, mean error was higher in all cases. Considering the additional effect of GPR, RMSE was approximately the same and mean error was slightly higher in all cases. Coverage was highest for Model A and second-highest for Model F in all scenarios. Note that it is possible for the ensemble of draws to have lower coverage than an individual model, for example, when the individual model's uncertainty interval fully encompasses that of the other component models. Coverage was otherwise roughly the same across models and scenarios, ranging from 0.591 (Model E in the full data scenario) to 0.672 (Model A in the "Missing years 1990 to 1999" scenario).

3.3.4 Observation 1: In first stage predictions, the ensemble model (E) had lower RMSE than any individual model.

Predictions for Bulgaria are an example of the overall result (across all locations) that the ensemble performed better in terms of RMSE than any individual model. Figure 3.2 shows

that variation in the model predictions for Bulgaria are approximately centered around the underlying true curve. This observation is especially apparent in years after 2000 where model-specific prediction variation is greater. Because the ensemble model (Model E) is an average of the four candidate models, the ensemble predicts well in this case. The four candidate models in these and subsequent figures are Model A: "lme4_default_v1", Model B: "nlme_v1", Model C: "spline_cascade_v1", and Model D: "xgbootstrap_v1". The red line represents truth, and the various model predictions are represented by lines of different patterns. The left column is the "Full data" scenario, the right column is the "Missing years after 2000" scenario, and the rows are different random realizations of the data generating mechanism. The figure shows only three of the five random realizations from the simulation.

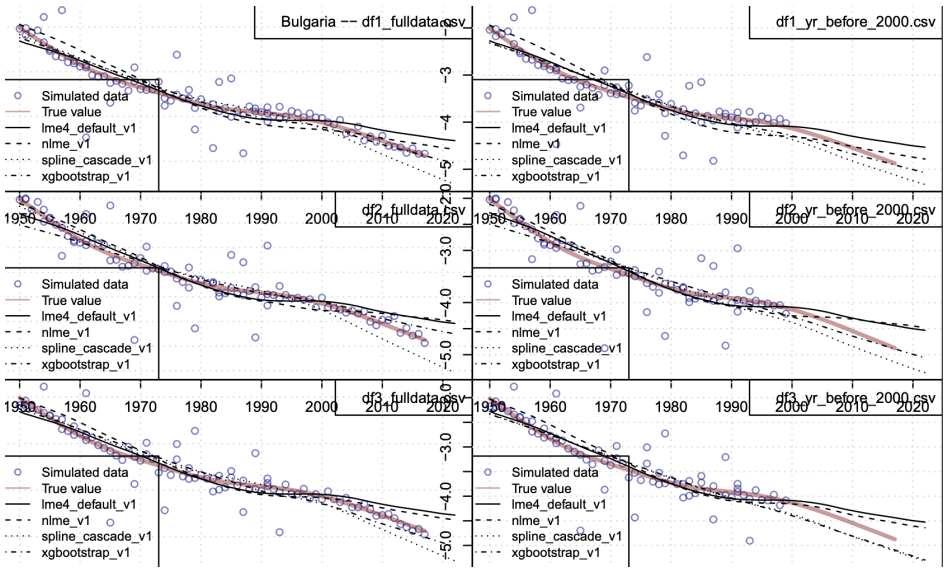

Figure 3.2: Under-5 mortality predictions for Bulgaria, three random realizations of scenarios with full data and data missing years 2000 and later

3.3.5 Observation 2: In first stage predictions, XGBoost did best on average among in-sample populations, but sometimes behaved erratically for out-of-sample populations.

The aforementioned Figure 3.2 is also an example of the overall result that XGBoost tends to outperform other models among in-sample populations. Where data are plentiful, the alternating dotted line labeled "xgbootstrap_v1" appears generally closer to the true curve compared to other models.

Where data were less plentiful, XGBoost sometimes behaved erratically across different random realizations of the data generating mechanism. Figure 3.3, for example, shows the various models' attempts to extrapolate trends into the future (right column). The plot in the first row (right column) shows a steep drop in predicted log-scale under-5 mortality during years 2004 to 2008 according to "xgbootstrap_v1". The plot in the second row shows essentially a linear decline (no drop), and the plot in the third row has a shallower drop. This occurred in spite of several measures taken to prevent overfitting, including monotonicity constraints, interaction constraints, hyperparameter optimization and bootstrapping. The covariate values for Timor-Leste shed some light on why this set of predictions might be unusual; the lag-distributed income (LDI) variable has a sharp increase during years 2004 to 2008. XGBoost fit on different random realizations of the same data generating mechanism, and with the same covariate values, handled the sudden LDI increase very differently. We discuss this observation in greater detail in the Discussion section.

3.3.6 Observation 3: Bayesian spline cascades outperformed other individual models for out-of-sample populations.

Figure 3.5 shows predictions for Comoros from the four candidate models. Extrapolations into the future (right column) are closest to the true curve for the spline cascade model (Model C), which is consistent with low RMSE and very low mean error for such extrapolations in the overall results.

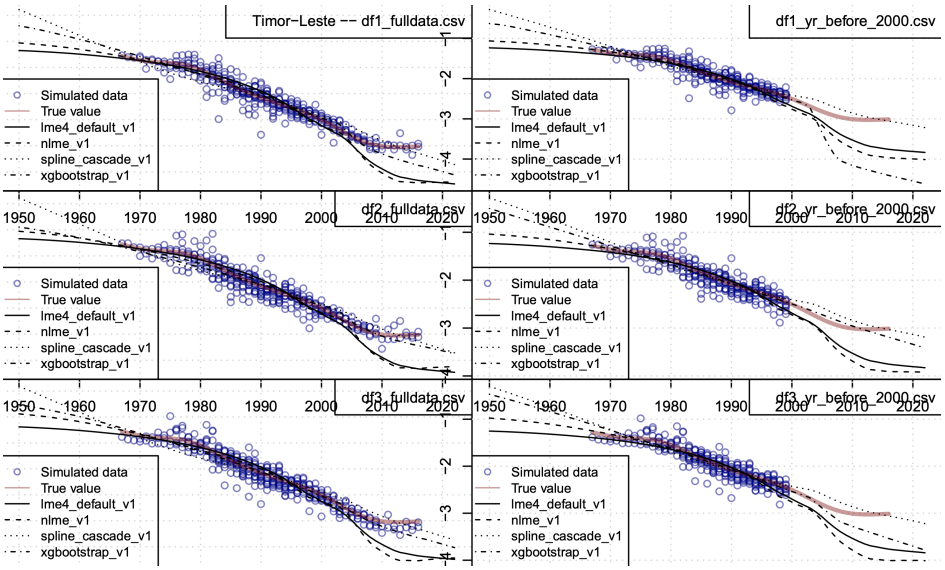

Figure 3.3: Under-5 mortality predictions for Timor-Leste, three random realizations of scenarios with full data and data missing years 2000 and later

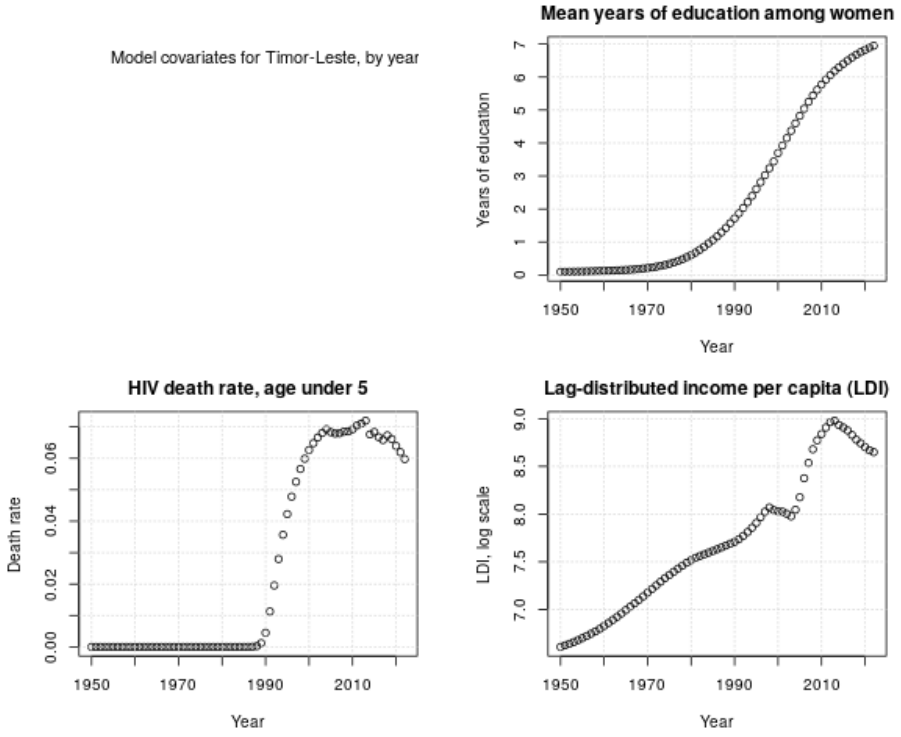

Figure 3.4: Covariate values for Timor-Leste over time, under-5 child mortality model

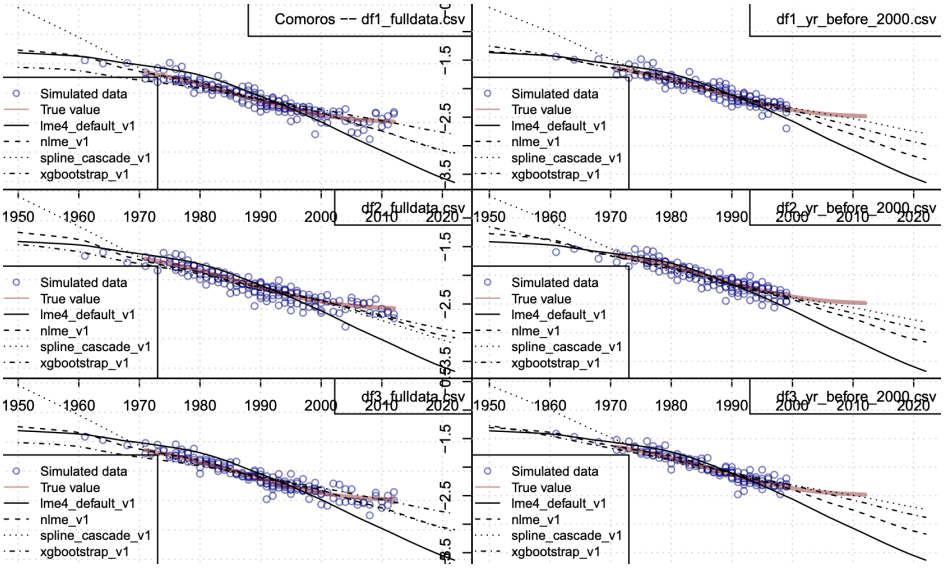

Figure 3.5: Under-5 mortality predictions for Comoros, three random realizations of scenarios with full data and data missing in years 2000 and later

3.3.7 Observation 4: ST smoothing reduced the RMSE of predictions in all scenarios; GPR did to a lesser degree.

Figure 3.6 shows an example from Guatemala of how smoothing residuals by geography and time ("space-time" smoothing, or "ST" in the plot) improves the model fit to the data. Whereas the purpose of the first-stage model is to capture general relationships between covariates and the outcome, the purpose of the ST step is to make predictions more sensitive to the characteristics of the particular population. The ST fit to the data is clearly an improvement over the first-stage model. GPR had relatively little additional benefit.

The ST step does not always bring the first-stage prediction back to the data, however. Figure 3.7 shows an example from the Netherlands where the first-stage model slightly underestimates the true curve in years after 2000, but the ST and GPR stages pull the prediction upward and farther away from the data. The pattern is similar for other countries that include only one data source as a data input to the model. Table 3.3 quantifies this observation by disaggregating performance metrics for the "Full data" scenario by whether a location

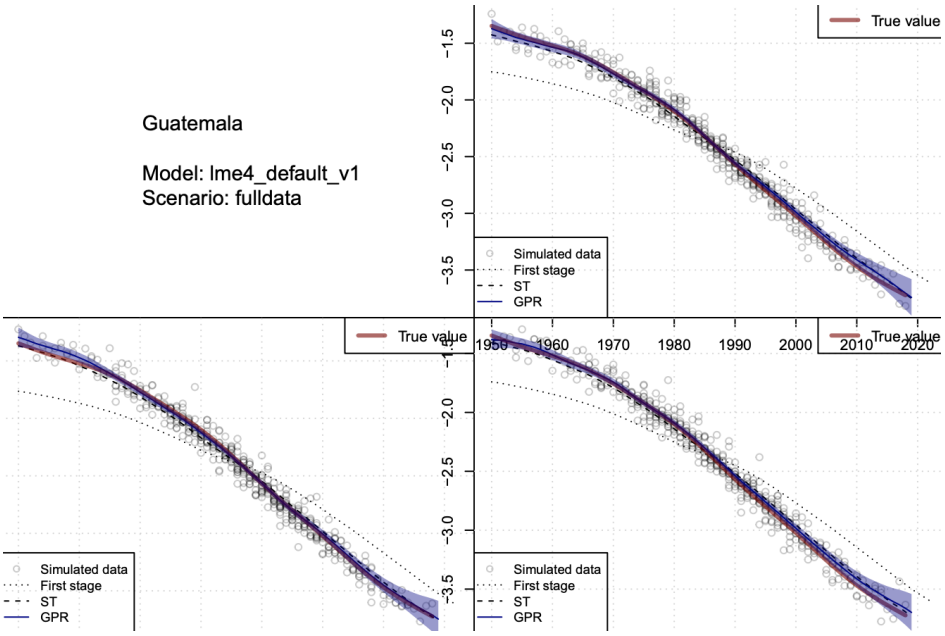

Figure 3.6: Under-5 mortality predictions for Guatemala, linear mixed effects model (Model A); three random realizations of scenarios with full data

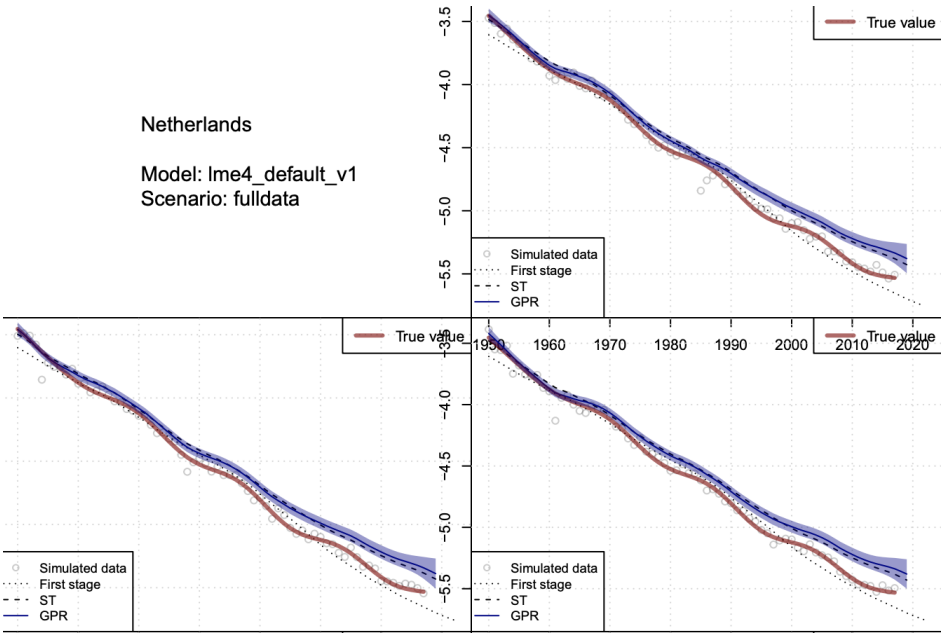

Figure 3.7: Under-5 mortality predictions for the Netherlands, linear mixed effects model (Model A); three random realizations of scenarios with full data

had one single data source, which was always a vital registration system, or multiple data sources. The countries with one source were: Andorra, Antigua and Barbuda, Australia, Austria, Barbados, Canada, Croatia, Czech Republic, Denmark, Dominica, Finland, France, Germany, Greece, Grenada, Iceland, Ireland, Israel, Italy, Japan, Lithuania, Luxembourg, Malta, Mauritius, Monaco, Netherlands, New Zealand, Norway, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, San Marino, Singapore, Slovakia, Slovenia, Spain, Sweden, Taiwan, The Bahamas, UK and USA. We found that ST-GPR had higher RMSE, higher mean error and lower coverage for these countries. Two explanations are likely. First, the higher quality of data from vital registration systems might be underappreciated in the ST step because residuals are not weighted according to data variance. Second, the hyperparameter governing the degree of spatial smoothing could be too high, in contrast to the GBD under-5 mortality model which essentially removes location-based smoothing for countries with VR data. Vital registration systems are more common in high-income countries where under-5 mortality is low, so spatial smoothing could lead to an upward bias. We discuss potential explanations for this finding further in the Discussion.

		Locations with multiple data sources			Locations with one data source (VR)		
Scenario	Model	RMSE	Mean error	Coverage	RMSE	Mean error	Coverage
Full data	Model A: Linear ME	0.1059	0.0418	0.724	0.1612	0.0927	0.408
Full data	Model B: Nonlinear ME	0.1064	0.0402	0.670	0.1614	0.0911	0.382
Full data	Model C: Spline cascade	0.1039	0.0415	0.687	0.1597	0.0918	0.382
Full data	Model D: XGBoost	0.1045	0.0416	0.666	0.1591	0.0910	0.363
Full data	Model E: Ensemble	0.1030	0.0409	0.665	0.1600	0.0911	0.364

Table 3.3: Performance metrics in locations with one data source (always a vital registration system) versus multiple data sources, average of five random realizations

3.3.8 Observation 5: Uncertainty intervals were too narrow but generally improved after making an ensemble of posterior draws.

We found that 95 percent uncertainty intervals included the true underlying value approximately 60 to 65 percent of the time in data rich prediction populations (Table 3.2) and

less often when extrapolating forward in time or predicting for locations with no data (Table 3.1). In the Discussion section we discuss possible explanations, including the potential effect of the bias noted in Observation 4. Even when mean predictions are unbiased on average, however, uncertainty intervals may be too narrow when not all sources of uncertainty are not incorporated into the model. For that reason, with existing model outputs, we assessed a new ensemble model (Model F) that takes 250 draws from each of the posterior GPR distributions of Model A, Model B, Model C and Model D. Performance metrics are given in Table 3.2 and Table 3.1.

Compared to GPR results in Table 3.1 and Table 3.2 for the other ensemble model (Model E) – taking the mean of first-stage predictions from the four candidate models and passing the result to ST-GPR – Model F had comparable RMSE and mean error, and higher coverage in every scenario. Coverage was especially improved for out-of-sample prediction populations ("Missing years 1990 to 1999": 0.704 versus 0.610, "Missing years 2000 and later": 0.656 versus 0.397, and "Missing 20 percent of locations": 0.911 versus 0.378).

3.4 Discussion

3.4.1 Overview

In this validation study, we evaluated several approaches for estimating levels and trends of under-5 mortality for every country. With such a complete historical record, data about under-5 mortality provide a robust empirical basis for validating descriptive epidemiological models. The main findings are consistent with the state of the science for predictive modeling: ensemble modeling improved both point prediction and uncertainty quantification, machine learning performed better than alternatives in data rich settings, and Bayesian models performed better than alternatives in data sparse settings. The simulation also provided insight into the Spatio-temporal Gaussian Process Regression (ST-GPR) model and potential areas for improvement. While ST-GPR improves point estimation for both data rich and data sparse settings in terms of Root Mean Square Error, there appears to be some

upward bias resulting from the space-time (ST) smoothing step. This bias is likely due to excessive spatial smoothing among countries with vital registration systems, a characteristic of this particular ST-GPR model that is not present in the current Global Burden of Disease under-5 mortality model.

3.4.2 Ensemble modeling and model uncertainty

Ensemble modeling is "a technique using multiple learning algorithms, or multiple statistical models, and combining them to improve estimates and predictive performance" [5]. Theoretical research has examined why ensemble models seem to have a comparative advantage over individual models [31; 32]. The ensembling methods used in the present study were relatively simple. Model E was the average of point estimates from Models A, B, C and D. Model F took 250 draws from the posterior distributions of Models A, B, C and D to obtain an overall mean estimate and uncertainty intervals. Among first stage predictions, Model E proved more effective than individual models in all scenarios for both in-sample and out-of-sample populations. The comparative advantage was lessened after passing these predictions to ST-GPR; one model in one scenario had better performance on RMSE, the spline cascade model in the task of extrapolating forward in time. Given the gains in performance observed in this study using a simple method, further research is needed to understand how more sophisticated methods of forming the ensemble might perform better, for example, weighting the models according to a cross-validated performance metric as suggested by Bannick and colleagues [5].

Model F incorporated model uncertainty into the predictions, resulting in uncertainty interval coverage closer to the optimal value compared to most individual models. The concept of model uncertainty refers to a degree of arbitrariness in the selection of one among many valid estimation procedures [24]. These might represent different yet equally valid sets of covariates, smoothing hyperparameters in a non-linear model, priors in a Bayesian analysis, or completely different modeling strategies as in this analysis. Even when a single model is chosen on principled grounds, model uncertainty is unaccounted for if other valid methods

would have yielded different results. Ensemble modeling allows for the quantification of such model uncertainty. Model F improved coverage the most in out-of-sample populations, likely because model predictions diverged the most under those conditions. The simulation also showed the counter-intuitive result that taking an ensemble of posterior draws does not necessarily increase coverage. If the uncertainty interval of an individual model completely encompasses those of the other component models, for example, the ensemble's uncertainty interval can be smaller in comparison to the individual model.

3.4.3 Machine learning

Machine learning (ML) algorithms, such as the XGBoost algorithm used in this study, are known as *big data* methods. ML has performed well in many settings [33; 34]. In the health field, ML has been used to improve clinical diagnoses [35], prediction of adverse health outcomes [36], health service delivery [37] and drug discovery [38] among other areas. Despite the success of machine learning in general, understanding the limitations of a specific algorithm for a specific use case is important. For example, one limitation of XGBoost is that it cannot extrapolate outside the domain of the feature space. For a single-covariate model with an increasing trend, for instance, XGBoost predictions will be constant for covariate values above the maximum value. This is a problem for descriptive epidemiological models that aim to make comprehensive global predictions. Data exist only for some locations in some years, so extrapolation beyond the existing data is required. We parameterized the XGBoost model as an annualized-difference model for this reason, to aid in the task of extrapolating forward and backward in time. While beneficial, our approach did not fundamentally solve the problem as time is not the only variable requiring extrapolation. Given the diversity and complexity of ML algorithms and their corresponding limitations, it is perhaps not surprising that many models have not lived up to expectations in real-world applications [39].

The problem with the use of big data methods in descriptive epidemiology is the need to predict for populations with little or no data. Big data methods require that data is, indeed,

big for the populations of interest. A distinction is required here. Big data methods – XGBoost, deep learning and others – achieve unparalleled results in terms of out-of-sample predictive validity when data are plentiful. The term "out-of-sample predictive validity" refers to the ability to guess what the value of a previously unseen observation is, based only on the observation's covariate values. When the data are fully representative of the population of interest, big data methods can take a dataset of millions of observations, separate the information content of the data from its idiosyncracies, and make out-of-sample predictions based on general patterns better than any other available method. Where big data methods might struggle, however, is taking a dataset with observations predominantly from high-income countries and making predictions for a low-income country. Predicting out-of-sample is not the same thing as predicting for data sparse areas of the feature space. In this sense, for descriptive epidemiology, the "big" aspect of big data methods might be considered an unfortunate requirement rather than a desirable attribute.

What is needed for making accurate predictions in data sparse areas of the feature space? The model needs to be supplemented with prior information, or information not included in the data itself. Often prior information is implemented as a particular constraint on the flexibility of the model. Again, a distinction is required here. Many big data methods have hyperparameters that control the degree of the model's flexibility, and these hyperparameters can be optimized through cross-validation. In the present study, for example, we optimized XGBoost's maximum tree depth and number of iterations through five-fold cross-validation in an attempt to minimize overfitting. This does not solve the problem, however, because while hyperparameter optimization does decrease the potential for overfitting in general, cross-validation optimizes with respect to the observations that already exist in the dataset. Data sparse populations are still relatively underprioritized. The additional information needs to come from outside the dataset, which is to say that the modeler needs to make a reasoned argument for constraining the model in a particular way. We implemented such constraints for XGBoost in the present analysis. The effects of income and education covariates on under-5 mortality were constrained to be negative, and the effect of the HIV death rate

covariate was constrained to be positive. These are established relationships in the literature on under-5 mortality [1]. We also implemented interaction constraints that forced the model to use regional covariate patterns of covariate effects rather than, for example, having a country-specific effect of income. This was a judgment we made based on the expectation of little heterogeneity of covariate effects within a region.

The result was that, on average, XGBoost had the overall best performance in data rich settings (see Table 3.2) and worse performance in data sparse settings relative to other models (see Table 3.1). We noted in Observation 2 that XGBoost sometimes behaved erratically. In Timor Leste, the effect of a sudden increase in lag-distributed income was handled very differently by XGBoost models fit on different random realizations of the same data generating process (see Figure 3.3). This occurred in spite of optimizing hyperparameters through cross-validation, encoding prior information into the structure of the model, and attempting to stabilize estimates by averaging over 30 bootstraps [40]. In principle, more constraints could be added to the model in order to stabilize the estimates further. In practice, it can be difficult to know when the implemented constraints are not adequate, especially outside of a simulation environment. If decision-makers are to make consequential choices based on a model's results, they deserve an explanation for why a prediction is a certain way. And, unfortunately, explainability methods like "SHapley Additive exPlanations" (SHAP) [41] do not help in this case. It is not sufficient, for example, to say that the SHAP values for lag-distributed income were large for one random realization and small for another. The difference is, by definition, random.

In summary, big data methods are not designed to predict well for data sparse areas of the feature space. The flexibility that is so valuable for data rich settings leads to instability in data sparse settings, a behavior that is problematic for models in descriptive epidemiology where extrapolation outside the data is usually required. For this reason, we hesitate to recommend big data methods like XGBoost when predicting for a comprehensive set of global locations despite performing better than alternatives in data rich settings. An irony of global health is that populations with the worst health problems also tend to have the worst

data. To the extent that population health information makes a real difference, selecting predictive models that deprioritize data sparse populations is likely to exacerbate health inequalities. As we describe in Chapter 2, for epidemiological datasets, flexible methods that let the data speak freely are perhaps better suited for tasks like exploratory analysis and hypothesis generation that do not require extrapolation to data sparse populations. Questions like "Why is the prediction like that?" can then be considered a prompt for more in-depth research on the topic.

3.4.4 Bayesian modeling

Bayesian modeling is well-suited to the task of predicting for data sparse areas of the feature space. Predictions must be informed by something, and in the absence of data the only alternative is prior information provided by the modeler. In the most general sense, any assumption or structure that constitutes the functional form of a statistical model can be considered a prior. Linear regression, for example, assumes among other things that the effect of a covariate is linear and errors are Gaussian distributed. These are important attributes of the model that affect the resulting predictions, but are not themselves derived from the data. Bayesian modeling extends this idea by allowing modelers to provide prior information about estimated coefficients. If we have reason to believe that the slope of the linear effect should be within 0.2 and 0.4, for example, we can communicate this to the model as a Uniform $[0.2, 0.4]$ prior in a Bayesian formulation of the model. The question remains, however, as to how a modeler should choose priors. A full answer to that question would require a diversion into epistemology and what counts as justified belief. For the purpose of the present study, we limit the discussion to empirical priors as implemented in the Bayesian spline cascade method.

In Chapter 1, we introduced a hierarchical Bayesian implementation of splines in which estimated coefficients from a model are passed as Gaussian priors to models fit on subsets of the data. The subsets in the present study are geographic subsets; coefficients estimated in the global model are passed as priors to region-specific models, and region to country. The

non-linear effect modeled as a spline was time. Thus the time trend estimated for a location results from a combination of the prior from the region-specific model and data in the particular location. Structuring the time effect as a geographical hierarchy is a modeling choice that, in itself, can be considered a prior that aids in predicting for data sparse populations. The actual coefficients passed as empirical priors further aid in this task. The result was that the spline cascade method had lower RMSE than all other individual models in predictions for data sparse populations. Considering ST-GPR outputs, the spline cascade method had lower RMSE for data sparse populations even compared to the ensemble. The improved performance was made possible by the inclusion of additional information not derived from the data, specifically that time trends should be correlated as a geographical hierarchy.

3.4.5 ST-GPR and uncertainty

ST-GPR reduced RMSE for all models in all scenarios among both data rich and data sparse populations. Most of these gains occurred in the space-time (ST) residual smoothing step, which borrows information from nearby observations more than observations farther away. The GPR step reduced RMSE somewhat in out-of-sample populations, but conferred no additional benefit in terms of RMSE in-sample populations. Distance metrics for ST smoothing are defined separately for age, for time and for sharing information across levels of the geographical hierarchy. We expected this process to improve predictions in data rich populations, and the finding that it improves prediction for data sparse populations further demonstrates its utility. Two observations require further investigation, however.

First, mean error was higher in ST-GPR outputs than it was for the corresponding first-stage predictions. With few exceptions, mean error was higher after ST smoothing compared to first stage predictions, and higher after GPR compared to ST smoothing. It is not immediately clear what causes this bias. One potential explanation is that residuals in the ST step are not weighted by inverse data variance in the process of smoothing. In this case, smoothing across time could lead to a plateauing effect at the extremes (i.e. recent years) in a way that is incongruent with the true strength of evidence for vital registration

systems, for example, which have low data variance. Figure 3.7 is an example. Alternatively, because vital registration systems are most common in high-income countries where under-5 mortality is low, and residuals are not weighted by data variance, the information sharing across locations leads to an upward bias in the places with vital registration systems. Finally, another potential explanation is an error in the code base for the GPR step, for example, that the mean of the posterior distribution might be calculated in linear space rather than the space in which the model is fit.

Second, coverage of the uncertainty interval was too low for all models in all scenarios. Incorporating model uncertainty (Model F) into the ensemble predictions generally improved coverage, as previously discussed, but coverage for Model F was still typically around 0.65. One potential explanation for the low coverage could be that the uncertainty interval is appropriately wide for an unbiased model, but the model is indeed biased. This is plausible as Table 3.3 showed that coverage was closer to 0.95 in locations where the prediction was less biased. A mechanism for this explanation could be that residuals in the ST smoothing step are not weighted by inverse data variance, but observations are weighted during the GPR step. For vital registration systems, for example, this would lead the uncertainty to be appropriately tight due to the use of inverse variance weights in the GPR step while still being subject to the bias from not using weights in ST smoothing. Another potential explanation is that data are used redundantly in the model. Observations are first used to generate the mean function as an input to GPR. Then they are used again in GPR itself. This explanation could be offset by the incorporation of non-sampling variance as an input to GPR in the form of marginal variance, but the topic may be worth further investigation in simulation. A third potential explanation is that inputs to the Matern covariance function of GPR might be inaccurate somehow. These include marginal variance, degree of differentiability and length scale, previously defined in Formula 3.13.

3.4.6 Limitations

This study had a number of limitations. First, the creation of a true underlying data generating mechanism for each location required making some distributional assumptions. These include that measurement error is Gaussian distributed in log space, random intercepts are Gaussian distributed in log space and that the mean function can be accurately represented as a thin-plate spline. We minimized the impact of these assumptions by using only location-years with at least two observations. We also visually inspected the spline fit for each location to verify that the time trend is reasonable. Second, while we attempted to roughly approximate the GBD's current method for under-5 mortality modeling, some differences do exist. The GBD model adjusts non-reference observations as part of the model fitting process, whereas we conducted the adjustment as a preliminary step. This was needed to ensure consistency across candidate models, so that differences can be attributed to the modeling strategy rather than the adjustment method. We also used the standard implementation of ST-GPR rather than the custom version used in GBD, which for example has differences in how the space-time smoothing weights are specified. The GBD model also varies the degree of spatial smoothing as a function of data density; data rich locations (i.e. ones with vital registration systems) share virtually no information across locations, while the implementation in the present study had a non-negligible degree of spatial smoothing applied equally to all locations. For this reason, conclusions drawn about Model B (non-linear mixed effects model) do not necessarily apply to the GBD model. Third, our implementation of XGBoost was customized to the modeling exercise. The choice of modeling annualized differences within surveys series, for example, was based on the observation that under-5 mortality sources tended to be repeated over many consecutive years, an observation that may not apply as readily to other modeled parameters. Fourth, other machine learning algorithms might have performed better than XGBoost but were not represented in this study. Conclusions drawn about machine leaning in general from this exercise should be updated in light of newer and potentially better alternatives. Fifth, the results of this study may not be generalizable to

other descriptive epidemiological parameters if the underlying data generating mechanism is substantially different. Sixth, using the unweighted average of model predictions is the simplest way to form an ensemble. Better methods exist, such as weighting components of the ensemble by an out-of-sample performance metric. However, the results of this study are sufficient to show that further exploration of ensemble methods is warranted for ST-GPR and similar descriptive epidemiological models.

3.4.7 Conclusion

Model validation is a key step in the process of developing descriptive epidemiological models, which aim to describe health loss in populations. Data on under-5 mortality provide a robust empirical basis for validating models designed to make comprehensive global predictions through time. The simulation confirmed several theoretical results from the predictive modeling literature: ensemble models perform better than individual models; machine learning performs best in data rich environments; Bayesian models shine in data sparse environments; and incorporating all sources of uncertainty, including model uncertainty, is important. The findings provided concrete examples for a fruitful discussion about the relative benefits of machine learning and Bayesian methods, in which we concluded that XGBoost may not be the most appropriate method for comprehensive global predictions given instability of estimates in out-of-sample populations. In contrast, the Bayesian spline cascade method performed well for out-of-sample populations. The study also found that ST-GPR substantially improved the precision of estimates relative to ground truth, although findings regarding higher mean error and low coverage of uncertainty intervals require further investigation.

3.5 Bibliography

[1] Wang, H. et al. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258):1160–1203, October 2020. Publisher: El-

sevier. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30977-6/abstract, https://doi.org/10.1016/S0140-6736(20)30977-6 doi:10.1016/S0140-6736(20)30977-6.

[2] Bates, D. et al. lme4: Linear Mixed-Effects Models using 'Eigen' and S4, June 2021. URL: https://CRAN.R-project.org/package=lme4.

[3] version), J.P.S. et al. nlme: Linear and Nonlinear Mixed Effects Models, February 2021. URL: https://CRAN.R-project.org/package=nlme.

[4] Murray, C.J.L. et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258):1223–1249, October 2020. Publisher: Elsevier. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30752-2/abstract, https://doi.org/10.1016/S0140-6736(20)30752-2 doi:10.1016/S0140-6736(20)30752-2.

[5] Bannick, M.S., McGaughey, M. and Flaxman, A.D. Ensemble modelling in descriptive epidemiology: burden of disease estimation. *International Journal of Epidemiology*, 49(6):2065–2073, December 2020. https://doi.org/10.1093/ije/dyz223 doi:10.1093/ije/dyz223.

[6] Sornette, D. et al. Algorithm for model validation: Theory and applications. *Proceedings of the National Academy of Sciences*, 104(16):6562–6567, April 2007. URL: https://www.pnas.org/content/104/16/6562, https://doi.org/10.1073/pnas.0611677104 doi:10.1073/pnas.0611677104.

[7] Krieger, N. *Epidemiology and the People's Health: Theory and Context*. Oxford University Press, March 2011. Google-Books-ID: Z59ciXRRaPsC.

[8] Principles of Epidemiology | Lesson 1 - Section 1, May 2020. URL: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html.

[9] Dieleman, J.L. and Haakenstad, A. The complexity of resource allocation for health. *The Lancet Global Health*, 3(1):e8–e9, January 2015. Publisher: Elsevier. URL: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(14)70373-0/abstract, https://doi.org/10.1016/S2214-109X(14)70373-0 doi:10.1016/S2214-109X(14)70373-0.

[10] Troeger, C.E. et al. Quantifying risks and interventions that have affected the burden of lower respiratory infections among children younger than 5 years: an analysis for the Global Burden of Disease Study 2017. *The Lancet Infectious Diseases*, 20(1):60–79, January 2020. Publisher: Elsevier. URL: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(19)30410-4/abstract, https://doi.org/10.1016/S1473-3099(19)30410-4 doi:10.1016/S1473-3099(19)30410-4.

[11] Gibbs, W.W. Bill Gates Views Good Data as Key to Global Health. URL: https://www.scientificamerican.com/article/bill-gates-interview-good-data-key-to-global-health/.

[12] Yoong, S.L. et al. Alignment of systematic reviews published in the Cochrane Database of Systematic Reviews and the Database of Abstracts and Reviews of Effectiveness with global burden-of-disease data: a bibliographic analysis. *J Epidemiol Community Health*, 69(7):708–714, July 2015. Publisher: BMJ Publishing Group Ltd Section: Review. URL: https://jech.bmj.com/content/69/7/708, https://doi.org/10.1136/jech-2014-205389 doi:10.1136/jech-2014-205389.

[13] Century, I.o.M.U.C.o.A.t.H.o.t.P.i.t.s. *Understanding Population Health and Its Determinants*. National Academies Press (US), 2002. Publication Title: The Future of the Public's Health in the 21st Century. URL: https://www.ncbi.nlm.nih.gov/books/NBK221225/.

[14] Kennedy, P. *A Guide to Econometrics. 6th edition.* Wiley-Blackwell, Malden, MA, 6th edition edition, February 2008.

[15] Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media, November 2013. Google-Books-ID: yPfZBwAAQBAJ.

[16] Dybowski, R. and Roberts, S.J. Confidence intervals and prediction intervals for feedforward neural networks. In Dybowski, R. and Gant, V., editors, *Clinical Applications of Artificial Neural Networks*, pages 298–326. Cambridge University Press, Cambridge, 2001. URL: https://www.cambridge.org/core/books/clinical-applications-of-artificial-neural-networks/confidence-intervals-and-prediction-intervals-for-feedforward-neural-networks/5DA7C939AEA8615D54D9771A6CEB9726, https://doi.org/10.1017/CBO9780511543494.013 doi:10.1017/CBO9780511543494.013.

[17] Boerma, T., Victora, C. and Abouzahr, C. Monitoring country progress and achievements by making global predictions: is the tail wagging the dog? *The Lancet*, 392(10147):607–609, August 2018. URL: http://www.sciencedirect.com/science/article/pii/S0140673618305865, https://doi.org/10.1016/S0140-6736(18)30586-5 doi:10.1016/S0140-6736(18)30586-5.

[18] Mathers, C.D. et al. Sensitivity and Uncertainty Analyses for Burden of Disease and Risk Factor Estimates. In Lopez, A.D. et al, editors, *Global Burden of Disease and Risk Factors*. World Bank, Washington (DC), 2006. URL: http://www.ncbi.nlm.nih.gov/books/NBK11802/.

[19] Chai, T. and Draxler, R. Root mean square error (RMSE) or mean absolute error (MAE)? *Geosci. Model Dev.*, 7, January 2014. https://doi.org/10.5194/gmdd-7-1525-2014 doi:10.5194/gmdd-7-1525-2014.

[20] Willmott, C.J. and Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, December 2005. URL: https://www.int-res.com/abstracts/cr/v30/n1/p79-82/, https://doi.org/10.3354/cr030079 doi:10.3354/cr030079.

[21] Marquaridt, D.W. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. 1970. https://doi.org/10.1080/00401706.1970.10488699 doi:10.1080/00401706.1970.10488699.

[22] Bethlehem, J. *Applied Survey Methods: A Statistical Perspective*. Wiley, Hoboken, N.J, 1 edition edition, July 2009.

[23] Durbin, J. Errors in Variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22(1/3):23–32, 1954. Publisher: [International Statistical Institute (ISI), Wiley]. URL: https://www.jstor.org/stable/1401917, https://doi.org/10.2307/1401917 doi:10.2307/1401917.

[24] Raftery, A.E., Madigan, D. and Hoeting, J.A. Model Selection and Accounting for Model Uncertainty in Linear Regression Models. 2007.

[25] Amaya, A., Biemer, P.P. and Kinyon, D. Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, 8(1):89–119, February 2020. https://doi.org/10.1093/jssam/smz056 doi:10.1093/jssam/smz056.

[26] Higgins, J.P.T., Thompson, S.G. and Spiegelhalter, D.J. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, January 2009. URL: http://doi.wiley.com/10.1111/j.1467-985X.2008.00552.x, https://doi.org/10.1111/j.1467-985X.2008.00552.x doi:10.1111/j.1467-985X.2008.00552.x.

[27] Foreman, K.J. et al. Modeling causes of death: an integrated approach using CODEm. *Population Health Metrics*, 10:1, January 2012. URL: https://www.

ncbi.nlm.nih.gov/pmc/articles/PMC3315398/, https://doi.org/10.1186/1478-7954-10-1 doi:10.1186/1478-7954-10-1.

[28] Flaxman, A.D., Vos, T. and Murray, C.J.L. *An Integrative Metaregression Framework for Descriptive Epidemiology*. University of Washington Press, 2015. Google-Books-ID: 2676pwAACAAJ.

[29] Wood, S. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, June 2021. URL: https://CRAN.R-project.org/package=mgcv.

[30] Pargent, F. et al. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *arXiv:2104.00629 [cs, stat]*, April 2021. arXiv: 2104.00629. URL: http://arxiv.org/abs/2104.00629.

[31] Laan, M.J.v.d., Polley, E.C. and Hubbard, A.E. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), September 2007. Publisher: De Gruyter Section: Statistical Applications in Genetics and Molecular Biology. URL: https://www-degruyter-com.offcampus.lib.washington.edu/document/doi/10.2202/1544-6115.1309/html, https://doi.org/10.2202/1544-6115.1309 doi:10.2202/1544-6115.1309.

[32] Laan, M.J.v.d. and Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York, August 2013.

[33] Schmidt, J. et al. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, August 2019. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Condensed-matter physics;Electronic structure;Materials science;Metals and alloys;Semiconductors Subject_term_id: condensed-matter-physics;electronic-structure;materials-science;metals-and-alloys;semiconductors. URL: https:

//www.nature.com/articles/s41524-019-0221-0, https://doi.org/10.1038/s41524-019-0221-0 doi:10.1038/s41524-019-0221-0.

[34] Young, T. et al. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, 2018. https://doi.org/10.1109/MCI.2018.2840738 doi:10.1109/MCI.2018.2840738.

[35] Choy, G. et al. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*, 288(2):318–328, August 2018. Publisher: Radiological Society of North America. URL: https://pubs.rsna.org/doi/full/10.1148/radiol.2018171820, https://doi.org/10.1148/radiol.2018171820 doi:10.1148/radiol.2018171820.

[36] Subudhi, S. et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *npj Digital Medicine*, 4(1):1–7, May 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Epidemiology;Machine learning;Predictive medicine;Prognosis Subject_term_id: epidemiology;machine-learning;predictive-medicine;prognosis. URL: https://www.nature.com/articles/s41746-021-00456-x, https://doi.org/10.1038/s41746-021-00456-x doi:10.1038/s41746-021-00456-x.

[37] Rose, S. Intersections of machine learning and epidemiological methods for health services research. *International Journal of Epidemiology*, 49(6):1763–1770, December 2020. https://doi.org/10.1093/ije/dyaa035 doi:10.1093/ije/dyaa035.

[38] Zhang, L. et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 22(11):1680–1685, November 2017. URL: http://www.sciencedirect.com/science/article/pii/S1359644616304366, https://doi.org/10.1016/j.drudis.2017.08.010 doi:10.1016/j.drudis.2017.08.010.

[39] D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[40] Breiman, L. Stacked regressions. *Machine Learning*, 24(1):49–64, July 1996. https://doi.org/10.1007/BF00117832 doi:10.1007/BF00117832.

[41] Lundberg, S.M., Erion, G.G. and Lee, S.I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]*, February 2018. arXiv: 1802.03888. URL: http://arxiv.org/abs/1802.03888.