Title

Innovations in machine learning: interval latent variables, causal exposure mixtures, and clinical predictive modeling

Author

Kennedy, Chris J

Publication Date

2020

Peer reviewed|Thesis/dissertation

Innovations in machine learning: interval latent variables, causal exposure mixtures, and clinical predictive modeling

by

Chris J. Kennedy

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alan E. Hubbard, Chair
Professor Mark J. van der Laan
Professor Mark R. Wilson

Fall 2020

Innovations in machine learning: interval latent variables, causal exposure mixtures, and clinical predictive modeling

Abstract

Innovations in machine learning: interval latent variables, causal exposure mixtures, and clinical predictive modeling

by

Chris J. Kennedy

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan E. Hubbard, Chair

This study presents three projects that build on machine learning techniques to propose new tools for scientific discovery: 1) constructing and measuring latent variables at scale by combining faceted Rasch measurement with supervised deep learning, 2) translating the problem of exposure mixtures (i.e. projecting multiple treatment variables onto a continuous summary measure) into a data-adaptive parameter within the targeted learning causal inference framework, and proposing an estimation algorithm, and 3) combining multiple techniques to improve the predictive accuracy and interpretability of clinical prediction models developed within the context of electronic health records.

The first project (Chapter 2) develops a general methodology to combine faceted Rasch measurement, a theoretically optimal form of item response theory, with supervised deep learning to construct and measure arbitrary interval variables. Rasch measurement theory is a method to create interval latent variables that are not directly observed, but can be approximated by collecting data on a set of components that are believed to contain information that indicates where an observation falls on that latent spectrum. The faceted version of Rasch modeling extends the method to rater-mediated assessments, which are essentially what most deep learning projects entail when they rely on human labelers to generate a training dataset. Bringing the formal tools of item response theory to machine learning offers a host of benefits, including reducing survey interpretation bias in the labeled data and upgrading the target variable from a dichotomous or ordinal structure to a continuous, interval variable, increasing precision. Perhaps most importantly, Rasch measurement theory provides a structure for gradual iteration based on the interplay between theorization and empirical testing, without which the field of machine learning has been forced to develop ad hoc solutions.

I realized in the course of the project that item response theory lent itself to a natural integration with a deep learning-based estimator for applying the constructed variable to

new data. While the deep learning model could be designed to predict the interval variable directly, as would be standard practice, an alternative architecture became visible: train the deep learning estimator to predict the individual scale components (items) from the labeling instrument, then apply the item response theory transform to those components in an offline fashion. Such an architecture leads to a new form of model explanation, because unlike standard neural models where the final dense layers are randomly initialized and generate their own internal latent variables that predict the final score, in our system those final latent variables were directly proposed via theorization and we had labeled data for which the neural optimization could provide supervised feedback. Predicting each item as separate outcomes in a single neural network entails a "multitask" architecture: the system simultaneously optimizes its predictive accuracy for each task, i.e. the predicted rating on each item. Multitask architectures are believed to offer efficiency gains because correlation between tasks allows information about one task (item rating) to also inform the model's prediction on other related tasks. And in an item response theory model the items would generally be highly correlated due to their coordinated measurement of an underlying latent variable. Predicting the rating on each item offers another twist: those item ratings are ordinal variables, so if we incorporate that scientific knowledge in our estimator we have the potential to gain efficiency and generate predictions that are consistent with that ordinal structure (i.e. predicted probabilities are unimodal across the possible ratings).

The second project (Chapter 3) examines the problem of estimating exposure mixtures, which are collections of treatment variables for which we seek to examine joint effects on a given outcome variable (e.g. disease state). Taking inspiration from the parametric method of weighted quantile sum regression, we recast the problem of exposure mixture estimation as a data-adaptive statistical parameter within the targeted learning causal inference framework. That allowed the establishment of an estimation procedure to nonparametrically project the vector of treatment variables onto a continuous latent variable that maximizes the joint relationship with the outcome, One could then evaluate causal parameters, such as treatment-specific means, on a held-out validation set using the cross-validated targeted maximum likelihood estimation procedure (CV-TMLE). Our method builds on earlier work with Alan Hubbard and Mark van der Laan as part of the varimpact algorithm, which is a data-adaptive method for causal variable importance that examines a single variable at a time. Through our mixture work we realized that an exposure mixture is a form of variable *set* importance, allowing subgroups of treatment variables to be ranked based on their combined mixture's estimated impact on the outcome variable.

The final project (Chapter 4), seeks to provide a guide to the development of high-quality clinical prediction models based on electronic health record data. Through the process of creating a risk prediction model for future heart attacks, we propose improved methods for many steps, including generalized low-rank models for missing data imputation, penalized histogramming to manage the cardinality of imputed covariates, nested Super-Learner ensembling for interpretable hyperparameter optimization, accumulated local effect

plots for model explanation, and the index of prediction accuracy as a general performance metric combining discrimination and calibration. The code for this project has been translated to a public dataset as part of a short course for training purposes, and is available at `https://github.com/ck37/Predictive-Modeling-in-R`.

To Katie, the latent variable of my heart

Contents

List of Figures

List of Tables

Acknowledgments

The wandering road to my PhD was paved with the collective action of a community of supporters. My first thanks go out to my advisor Alan, who has benevolently aided my journey through the years. He initially hired me as a political science PhD student who for some reason wanted to work on medical machine learning back in 2016. With Alan's blessing it then became possible for me to switch over to the biostatistics PhD program to focus on causal inference and machine learning, and I don't know if many other advisors would have been open to such an unorthodox path. Between finding me funding, research projects, and collaborations, I owe a vast swathe of my career direction to Alan's selfless leadership.

Several other advisors deserve heartful thanks for their support. Mark van der Laan has been an inspiration to work with, and was a major motivator for me choosing Berkeley for my PhD. Catherine Metayer originally tasked me with developing an exposure mixtures method back in 2017, and has patiently cultivated my work on that project. I hope to one day run a lab half as well as Catherine has demonstrated with the Integrative Cancer Research Group. Mark Wilson opened my eyes to Rasch measurement theory, inspiring an essential rigor that the hate speech project would otherwise have lacked. Claudia von Vacano has been gracious in her support of my work through D-Lab, a gallant leader on many levels, and instrumental in the creation & continuation of the hate speech project. Mary Reed enthusiastically supported my meandering work on clinical prediction modeling at Kaiser Permanente; her penchant for linking research to direct impact on care remains a crucial guiding star. My gratitude goes to Chris Mann for facilitating & mentoring my move from industry to a PhD back in 2013.

Additional thanks are warranted for my fellow students, colleagues, and collaborators in biostatistics, political science, D-Lab, the Berkeley Institute for Data Science, the Integrative Cancer Research Group, the School of Public Health, Bakar Institute, and the Kaiser Permanente Division of Research. Without their camaraderie it would have been a far more difficult and depleting graduate experience.

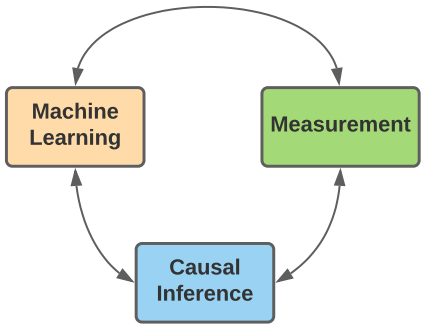
I am grateful to my parents for teaching me a love of learning and for supporting me unconditionally. I thank my beagle children Edith and Walter, whose affection has brightened many a tiring day. Finally to Katie, who makes it all worthwhile.

Chapter 1

The scientific triumvirate: machine learning, causal inference, and measurement

In this dissertation I describe three interconnected projects that contribute to new methodological developments, all incorporating machine learning. These projects represent the focus of my research activity over the past 3-4 years. Through these projects, I have been led to a perspective on empirical scientific reasoning that is integrated around three primary pillars of activity: machine learning, measurement, and causal inference. I call these fields the *scientific triumvirate*.

Figure 1.1. The scientific triumvirate: machine learning, measurement, and causal inference



1.1 Machine learning

By machine learning I largely mean "prediction", although with the implication that a wide variety of estimation algorithms will be employed. Prediction is helpful in innumerable ways.

It is so severely pervasive in our daily lives that we hardly notice it. We check the morning weather to calibrate our expectation of the temperature, precipitation, and possibly air quality for the remainder of the day. We shop for consumer goods based on the expectation of future enjoyment. We delay our commutes or leave early in order to compensate for additional traffic during rush hour. The task of prediction is one of learning a mapping from a set of predictor variables to a desired target or outcome variable.

Although prediction is the primary task, my focus on machine learning is intended to emphasize the benefit of complex estimators. The reliance on machine learning originates from a desire for scientific humility. In nearly any real-world application we do not know the true relationship between the observed outcome(s) of interest (e.g. a disease state or treatment variable) and observed predictor variables. If we acknowledge that ignorance we cannot in good faith trust the prediction of a single algorithm, such as a linear regression estimator, without evaluating its accuracy with data that we already have. Cross-validation becomes the crucial tool to (partially) alleviate this ignorance of functional form. By teaching our estimator with supervised examples on the training set, and evaluating its performance on a separate unseen test set, we can evaluate the correspondence between our estimator's prediction and new data. That evaluation then allows the ranking of competing estimators, obviating the necessity of assuming that a particular estimator perfectly represents the unknown relationship between the predictors and the outcome.

Once model evaluation through cross-validation is adapted, it is a small additional step to adopt ensemble learning. Ensemble learning is a style of supervised learning in which we allow our prediction algorithm, or estimator, to consist of an aggregation of one or more constituent algorithms, called "learners." Each learner consists of an arbitrary estimation algorithm, such as linear regression or random forests, combined with specific configuration settings (hyperparameters), and possibly additional subsetting of the available predictor variables (feature selection or screening) or data preprocessing (e.g. a coordinate transformation via principal component analysis or imputation of missing values with generalized low-rank models). Through cross-validation we can honestly estimate the predictive accuracy of each learner and select the version with the best performance (lowest loss). Ensemble learning takes that a step further by recognizing that even better performance might be achieved by combining the predictions of multiple individual learners, forming a team prediction that has the potential to make fewer errors than any individual member. The task of aggregating the predictions of the cross-validated learners is called *meta-learning*, and might be accomplished through a weighted average (convex combination) of individual learners that is designed to maximize predictive accuracy. We call such an algorithm to take the test-set predictions from cross-validation and create a final prediction estimator with meta-learning a *super learner* (M. J. van der Laan, Polley, et al. 2007), also known as stacking (Wolpert 1992). The process of developing, evaluating, and interpreting such ensemble models is the focus of Chapter 4, particularly in the clinical setting of electronic health records.

1.2 Causal inference

Despite the appeal and intuitive value of machine learning for prediction, in many use cases prediction is a placeholder or simplification of what would better be examined through the lens of causal inference. Causal inference represents the *sine qua non* of scientific pursuits: the estimation and testing of hypothesized treatment effects that build empirical understanding of mechanisms. In a causal inference setting one posits a counterfactual intervention, where the fundamental question is one of so-called *treatment-specific means*: if I treat or expose a group of units in a certain way, what is my best guess (point estimate) and range of expectation (confidence interval) for the average outcome of that group? From that foundation we can build out more complex target parameters, such as comparing two treatment-specific means (average treatment effect), examining subgroups for variation in that comparison of treatment-specific means (conditional average treatment effect), examining the combined effect of multiple interventions (mixtures), or considering changes to the treatment or exposure across multiple time points (dynamic treatment regime or rule-specific mean).

Problems that are initially framed as prediction tasks often have an underlying question centered around causal inference, but which may not be immediately apparent. In the clinical setting, for example, there is often a desire to predict patient risk of overall mortality or certain diseases. Certainly it is often useful to the patient and to the clinician to have an estimate of the likelihood of death or disease. But when one better understands the intended usage of the risk score, it can very often shine the light on a true desire for a causal inference analysis. The desire for patient risk estimates often stems from a presupposition that patients could be treated differently based on that prediction, e.g. low-risk patients might receive a less aggressive treatment plan compared to high-risk patients. If that is how the risk score is intended to be used, the question of scientific interest is more appropriately thought of as one of optimal individualized treatment. The clinician asks: given the data I have on this patient, which of the available treatments do I estimate will yield the best result? When framed in that way we can see that the risk score is only one of likely many patient characteristics that could influence our estimate of which treatment is most effective. The prediction paradigm makes the implicit assumption that the risk score is the most valuable differentiator of patient outcomes under different treatment strategies, whereas the causal inference paradigm seeks to directly estimate the expected outcome under each possible treatment and using all available predictor data.

Similar tensions of prediction versus causal inference arise in the course of political campaigns. On the prediction side one could create estimators that predict the likelihood of voting in an upcoming election, of supporting a certain candidate, or of agreeing with a certain message. But the true set of questions underlying those prediction tasks is one of *intervention*: which voters should the campaign attempt to motivate to vote with a limited budget, which message framing for a given voter would most increase support for the candidate, and which voters are not cost-effective targets for the campaign's available interventions? These questions are ones of causal inference because they involve estimating the

likely effect of actions that cannot be directly observed on every individual in advance.

In this work I focus on the targeted learning framework for causal inference (M. J. van der Laan and Rubin 2006; M. J. van der Laan and Rose 2011, 2018), which was a significant focus of my graduate training as well as a key reason that I chose Berkeley for my PhD. Targeted learning has several notable scientific features: 1) it is a double robust method, incorporating outcome regression estimation (covariate adjustment) and inverse propensity score weighting (IPTW), 2) it is designed to incorporate supervised machine learning for both of those estimation steps, avoiding unjustified assumptions about the relationship between variables, 3) it is effective for both randomized trials and observational studies, unlike many methods that are designed for one or the other, 4) it is agnostic as to the target parameter of scientific interest (i.e. we are not bound to an odds-ratio or hazard ratio as the desired scientific goal), 5) it makes efficient use of acquired data, giving confidence bounds around its best guess that are trustworthy. Although its mathematical complexity makes targeted learning a non-trivial framework to thoroughly understand, these features reward the user with a system that can effectively answer a vast variety of causal inference questions.

1.3 Measurement

The final, and possibly most unorthodox, element of my scientific outlook is the importance of measurement, particularly when combined with causal inference and machine learning. The field of measurement originates from psychometrics and educational statistics, which has been driven by the development of standardized testing such as the Intelligence Quotient (IQ) test, Scholastic Achievement Test (SAT), and the Graduate Record Examinations (GRE). Measurement theory is concerned with the development of exams, or survey instruments, that can consist of a series of questions, or items, and response options to those items. If the instrument was developed with adequate care, theorization, and testing, then analyzing the patterns of answers on those items can allow one to reliably estimate the value of each participant on an underlying variable that is never directly observed (i.e. a latent variable). In the case of a test of knowledge or skill, such as the GRE mathematics or verbal exam, the more questions that an examinee answers correctly, the greater the estimated latent skill in that subject area. Or the exam might instead be designed around a subject where answers are not right or wrong, but are simply indicative of a certain area or range on the latent variable, such as a personality trait or general attitude. Ultimately we use measurement to carefully design and build a ruler that provides a stable, numeric value for the latent variable which can be compared between objects, and to check the straightness of the ruler across different samples and subgroups.

Measurement becomes important in situations where we want a precise value for a variable, when we want to understand changes over time and across samples, when we want to be able to explicitly and thoroughly check for fairness or bias, and when third-party reviewers are evaluating subjects. For most variables in a scientific dataset it would be overkill to develop a multi-item instrument and apply a measurement model to combine those answers

to estimate the value of a theorized latent variable. Yet the outcome variable has a special status when conducting studies - changes in the value of the outcome variable are a central focus in randomized trials and observational studies. That outcome variable may represent an entire subfield of study, such as a disease. The outcome may be difficult to precisely define and be subject to disagreement, and especially when human raters are involved there may be worries about biased ratings that are related to demographic differences (say race or gender). And we may want to run meta-analyses to compare treatment effects across studies, with some effort to put each study's outcome variable on a similar scale. The more that we care about these issues for a variable, the more it becomes worthwhile to adopt a measurement outlook that provides specific tools to deal with these complications.

Somewhat surprisingly, measurement is not a common part of the toolkit of the typical statistician or data scientist. It does crop up here and there in certain areas though. In political science measurement is used to created ideological scales of politicians based on their voting records, an area of study called spatial modeling (referring to an geometric space of political ideology defined by 1 or 2 latent dimensions). In medicine, the skill of surgeons is evaluated by human review of videos that are rated on certain indicators. In randomized trials and survey research the primary instrument might contain multiple questions related to the outcome of interest, though they might be aggregated in an ad hoc manner using principal component analysis. I hope through the hate speech project (Chapter 2) and through future work that my research agenda can make a case for measurement to be more commonly integrated into the toolset of the average statistician or data scientist. I believe that the framework developed in M. Wilson (2004) in particular is likely to gain prominence as the value of measurement theory becomes better understood.

Among machine learning researchers in particular there is a compelling benefit to understanding the tools and techniques of measurement theory. As will be demonstrated in Chapter 2, when human workers create labeled datasets for machine learning models there is an opportunity to aggregate those labels using measurement theory in thereby reduce survey interpretation bias, improve reliability, and gain the capacity for more formal study of demographic bias. The centrality of human-labeled training data for machine learning models makes the benefit of measurement theory arguably even more pronounced than in causal inference research. Issues of bias & fairness, reliability, and reproducibility are in the zeitgeist of machine learning research, and measurement theory offers techniques backed by decades of research that can be leverage to address these issues. The fields of computer science and statistics stand to make much quicker improvements in their approaches to machine learning if they can avoid reinventing the wheel and instead directly integrate the substantial benefits of measurement theory.

1.4 Conclusion

In sum, I see the framework of causal inference, particularly targeted learning, as representing the most valuable, central role in my scientific research agenda as it effectively unifies obser-

vational and randomized studies, incorporates ensemble machine learning for estimation, and can take advantage of any precise outcome variables created through measurement. Chapter 2 describes the integration of measurement theory with machine learning (particularly deep learning) for the design of effective human-annotated datasets and debiased prediction algorithms, with application to hate speech. Chapter 3 proposes a new approach to exposure mixtures based on the targeted learning framework, with application to a case-control study of childhood leukemia. This work is inspired in part by my earlier work on variable importance estimation as a data-adaptive target parameter (Hubbard, C. J. Kennedy, et al. 2018). Chapter 4 promotes a focus on high-quality clinical prediction models with detailed approaches to missing value imputation, hyperparameter optimization, and interpretation. Each of these core chapters is based on co-authored work in progress, so I highly recommend that those later publications be used as the primary references rather than their current state as dissertation chapters.

Chapter 2

Integrating Rasch measurement theory with deep learning

2.1 Introduction

Across fields of knowledge (science, engineering, medicine, etc.) phenomena of interest are often labeled by humans as discrete, dichotomous variables: speech is toxic or not, an MRI scan shows cancer or is clear, a stoplight is red or green. Simple variables may be inherently binary or ordinal in nature. For example, an online advertisement might be clicked by a person viewing a webpage, or might not — a partial click between 0 and 1 does not make sense. But if a discrete variable represents a concept with complexity or granularity, those {0, 1} values may reflect the simplification of an underlying continuous spectrum or latent variable. Consider an online message that is rated for its sentiment using the discrete, ordered labels of "positive", "neutral", or "negative" (B. Liu 2015). Those labels can reasonably be viewed as summarizing what is ultimately a continuous, infinitely divisible spectrum ranging from extremely negative to extremely positive sentiment.

Physical quantities such as temperature and weight can be measured as interval variables where magnitudes are meaningful: we can subtract the measurements of two units and have an estimate of distance on the original scale. We can make factual statements such as "today's temperature is 5 degrees warmer than yesterday" or "I lost 10 pounds after dieting". The development of physical measurement systems required major, long-term investments in scientific and engineering that led to a theory of those physical systems (Chang 2004). Thanks to those fixed scales we can receive a weather report with specific temperatures provided, or stand on a scale and view our current weight. But in the fields of machine learning, natural language processing, and many other areas, current practice would provide a binary prediction. As a mental exercise, what if today's expected weather forecast were reported as two values: hot or cold? Classification models could only estimate $Pr(Weather = Hot \mid Data)$ and we might hear: "Today's weather forecast: expected hot, with 55% probability, but a 45% chance of being cold." It sounds absurd, but that is how we treat many variables in

science today, including our construct of interest in this paper: hate speech. "Classification" of discrete outcome variables has seemingly become synonymous with "supervised learning", so rarely are continuous variables analyzed. Our question becomes: how can we aspire to emulate physical scales like temperature, which we take for granted in our daily lives, and construct interval measurements for arbitrary variables?

A method to estimate a continuous spectrum for human-generated variables could be valuable for several reasons. If the variable is recorded in order to determine the application of interventions, a continuous measurement would allow many different thresholds to be defined with varied policy prescriptions. A binary outcome would support only two policy alternatives, and research based on the binary version would lose most of its underlying value if the implicit threshold of interest changed (Streiner 2002). When the construct of interest is an outcome variable, such as in a randomized trial or an observational study, having a continuous measurement will increase statistical power compared to a dichotomous or ordinal variable (Cohen 1983; Senn 2003). If the variable is a confounder in a causal inference analysis, a continuous measurement would more strongly reduce residual confounding compared to a discretized version of that variable (Royston et al. 2006). Similarly, if the variable is a pre-treatment adjustment variable in a randomized trial, the variable will more effectively improve precision if continuous rather than binary (Dawson et al. 2012).

In this work we propose a methodology to construct continuous, interval variables by combining two complementary techniques: many-facet Rasch measurement, a form of item response theory, and supervised deep learning. Rasch measurement theory involves the application of a probabilistic multilevel statistical model to create continuous, interval scales out of multiple survey questions, measured as binary or ordinal variables. When those survey items (or "components") are completed by human test-takers we need a way to automate the human ratings for unseen future data, otherwise we would need human test-takers to review any future observation in order to create its measurement on the scale. Supervised deep learning provides an automated approximation of the human ratings by learning to predict those ratings on arbitrary new observations, which also serves as a new form of model explanation. The deep learning model can alternatively predict the continuous scale directly and skip the intermediate step of predicting the ratings. Our methodology can be applied to any variable that can be theorized as a continuous spectrum and is based on a human ratings. The underlying source data is also general: it might be text, images, videos, time series, waveform, or audio. As long as the underlying data source is unstructured and reviewed by humans, the deep learning model has the potential to approximate those human ratings for automation purposes.

We apply this novel method to the measurement of hate speech, a social problem that has received extensive attention from policymakers, researchers, and firms. While existing attempts to *detect* hate speech use predictions derived from discrete labels, our method *measures* the hateful content of speech by scaling multiple labels of multiple components.

Our manuscript is structured as follows: we begin by providing background on the theoretical foundations of hate speech and previous machine learning approaches. We then describe our methodology, starting with the theoretical development of a hate speech spectrum,

which is operationalized as a survey instrument, the collection of social media comments and crowdsourced labeling process, use of faceted Rasch theory to statistically transform the labeled data into an interval variable, and deep learning model to predict that interval variable using only the text of the comment. We next present the results of our work, which consist of the Rasch scaling followed by the deep learning performance evaluation. We conclude with a discussion of current limitations and the many possible extensions of this method. We hope that this work can help spur wider usage of Rasch measurement theory to go beyond oversimplified discrete variables and instead develop continuous, interval variables across fields.

2.2 Background

Application to hate speech

Hate speech is immensely consequential in our current political and economic context. The accurate identification, measurement, and intervention of hate speech may prevent psychological harm, dissipate extremist groups and thereby prevent violent, even genocidal events, and may increase the quality of information that is being distributed online.

The harm in hate speech is significant. Within niche extremist communities, hate speech serves as a mechanism for recruitment and as a radicalizing influence, building cultures of hatred that can lead to hate crimes and terrorist violence (Tsesis 2002). In areas with long-term ethnic conflicts, hate speech may precipitate or otherwise foment mass genocidal violence (R. A. Wilson 2017).

Computationally identifying hate speech is a challenging task. First, there is no consensus on a systematic definition of hate speech (Sellars 2016). Second, the scale of communication and the different linguistic forms it takes necessitates algorithmic approaches to detection. Third, hate speech constitutes a small proportion of online communication (1% or less), making it difficult to create representative training datasets for algorithmic approaches. Fourth, human reviewers that generate the training data often do not agree on how to label training observations, even when given consistent definitions to evaluate speech. Finally, extremist groups often use sarcasm, ambiguity, and coded language to obscure the hateful nature of their communications, to evade moderation, and ultimately, to organize actions. The combined impact of this mixture of challenges is that we cannot currently measure hate speech in a consistent manner over time. This, in turn, makes it impossible to understand how durable hate speech is, or what the causes of hate speech are.

This article proposes a new approach that addresses these five challenges, integrating *Constructing Measures* & Rasch measurement theory with deep learning for natural language processing (NLP). We describe the development of a hate speech scale, grounding its theoretical development empirically through a reference set, estimating the scale on labeled comment data, correcting for the bias of individual reviewers, and testing that scale through IRT diagnostics. We show how IRT integrates naturally with a multitask deep learning

architecture that improves sample efficiency, increases accuracy, explains its reasoning, and provides more granular predictions compared to non-IRT methods.

Computational hate speech analysis

Machine learning for hate speech analysis has become a well-studied topic over the past decade. Early work (e.g. Warner et al. 2012) used simple, non-deep learning algorithms for estimation, such as naive Bayes, support vector machines, or linear regression. Text features were typically unigram TF-IDF scores, and sometimes additional syntactic features like part-of-speech tagging (Davidson et al. 2017). The field gradually transitioned to word embeddings, often pre-trained, where the best featurization approach was comment word embedding averages or paragraph embeddings (Djuric et al. 2015; Nobata et al. 2016). Those features could then be combined with gradient boosted decision trees or Bayesian additive regression trees.[1]

More recently there has been a shift to deep learning methods, beginning with Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) networks, that either use pre-trained word embeddings or learn their own embeddings from the raw text (Badjatiya et al. 2017; Zhang et al. 2018). Adding attention mechanisms to deep learning is beginning to be used in hate speech research (Founta et al. 2018), but their successor transformer units (Vaswani et al. 2017) have not yet been widely adopted for hate speech research.

Model-based transfer learning methods provided a further breakthrough (Ruder 2018), including best in class architectures OpenAI's GPT-2 (Radford et al. 2019), fast.ai's ULMFiT (Howard et al. 2018), and Google's BERT (Devlin et al. 2018). Those contextualized word representation methods have not yet been used in hate speech research but we expect them to be soon.

We summarize the comparison of related work on hate speech in Table 2.1.

2.3 Methods and Data

We use our method to measure hate speech, a complex social phenomenon that is difficult to define, measure, and detect at scale. Our methodological approach constructs a systematic definition and measurement of hate speech in social media text and creates a more efficient and less biased algorithm for detection. In the following sections, we describe the process of construct theorization using a reference set of comments, the measurement of sub-constructs using a labeling instrument, the crowdsourcing of labeling to a network of linked reviewers, the debiased scaling of labeled data using Rasch measurement theory, and the integration of the Rasch model into a novel deep learning architecture for debiased, explainable prediction.

[1]These were our best results from phase 1 of the current project, completed in October 2017.

Table 2.1. Comparison of related work. Outcome measurement granularity refers to the measurement model for the raw scores (i.e. cardinality). CTT = Classical Test Theory. IRT = Item Response Theory. Obs = Observations. SVM = Support Vector Machine. CNN = Convolutional Neural Network. LSTM = Long Short-Term Memory Neural Network.

Paper	Construct levels	Outcome measurement granularity	Measurement theory	Identity categories	Algorithms	Obs.	Platforms	Ratings per comment	Reviewer bias adjustment	Coded language annotation
Warner et al. 2012	2	2	CTT	1	Linear SVM	1,000	Yahoo! Groups	3	No	No
Davidson et al. 2017	3	3	CTT	1	Lasso, Ridge	24,802	Twitter	3	No	No
Del Vigna et al. 2017	3	3	CTT	3	LSTM	3,685 - 6,502	Facebook	1-3	No	No
Dixon et al. 2018	2	2	CTT	30	CNNs	127,820	Wikipedia	3?	No	No
Siegel et al. 2019	2	2	CTT	1	Naive Bayes	25,000	Twitter, Reddit	3	No	No
This study	8	32 - 48	Faceted Rasch IRT	50	USE, BERT, RoBERTa	50,000	YouTube, Twitter, Reddit	4	Yes	Yes

Construct theorization from a reference set

Hate speech has many definitions across academic disciplines, legal and regulatory doctrines, and in common vernacular. When attempting to systematically measure a social concept, "I know it when I see it" will vary greatly on who is doing the identification: life experience, familiarity with language, and historical context all vary across individuals, whether experts or laypeople. Difficult-to-measure phenomena like hate speech are an ideal application of our method, but require careful theorization and translation to measurement instruments.

We draw from the legal definition of hate crimes in the United States that protections against discriminatory actions targeting one of the following protected groups: race, religion, ethnicity, nationality, gender, sexual orientation, gender identity, and disability. In identifying groups within these broad categories, we include subjugated groups that have been discriminated against in the United States, as well as power-dominant groups who have not. Targeting of a group or an individual on the basis of their membership in a group is common to most definitions of hate speech (Sellars 2016). Not only do we adopt this convention, but we allow for intersectional or overlapping identities to be selected for further analysis. We consider intersectional identities and the possibility of compounding hate speech directed at an individual who belongs to multiple groups.

Speech can also lead to individual acts of violence and when targeted against a group, genocide and extermination. The "dangerous speech" framework ties the effects of hateful speech to actions that it can incite (Benesch et al. 2018). Dehumanization, such as radio broadcasts in Rwanda referring to the Tutsi people as cockroaches, is directly linked to later genocidal killing of that group. Incitement towards violence is a narrowly defined

concept under United States law, and the dangerous speech framework that we use takes a broader view of the link between cause and effect. Sellars (2016) points out that the accumulated affects of anti-Semitic or racist speech can have multi-generational impacts on well-being of individuals of group born long after hateful speech was original created. Given the complexities of these concepts, we focus on calls to individual violence or collective extermination, with the idea that these are the final step after expression of hate and deeming a group inferior or inhuman.

Theorized construct levels

Table 2.2 describes the eight levels of our theorized hate speech scale. The positive levels on the scale designate hate speech of increasing severity. Unlike many existing scales, our typology includes both neutral and positive identity speech, represented by 0 and positive values, respectively. Speech that does not address protected identity groups does not fit into our typology.

Following Stanton (2013), we place speech supporting the systematic killing of a specific group as the most severe form of hate speech. Viewing other types of hate speech as pathways to genocide, we pay special attention to individuals threats of violence and dehumanization that may justify violence.

Table 2.2. Theorized qualitative levels of the hate speech — counterspeech spectrum.

Level	Summary	Description
5	Genocide	Support for or intention of systematically killing all or a large number of a protected identity group
4	Violence	Threat or support of physical force or emotional abuse intended to hurt or kill members of a protected identity group
3	Dehumanization	Depriving a protected group of human-like qualities, such as comparison to an animal, insect, or disease
2	Hostility	Unfriendliness or opposition to a protected identity group, such as through slurs, profantiy, or insults
1	Bias	Inclination or preference against a protected identity group, including prejudice
0	Neutral	Descriptive or other non-harmful references to identity groups
-1	Supportive	Respectful, prideful, or other solidarity-based messaging about a protected identity group(s)
-2	Counterspeech	Response to hate speech that seeks to undermine its impact and standing

From a manual review of social media comments, we curated a *reference set*, a small corpus of example text for each conceptual level. We selected 10 comments to serve as examples of each of our theoretical levels, totalling 80 comments. These comments serve two purposes: to aid in construct development by providing clear examples between levels and as empirical scaffolding to link human raters in our crowdsourced measurement. We describe the importance of the reference set for construct theorization and measure development below and its importance for linking labelers in Section 2.3.

In concert with construct development using existing literature, our research team manually reviewed thousands of comments from our corpus, oversampling on comments with high hypothesis scores. We also selected reference set comments for each level that yielded a diversity of target groups, text length, and linguistic styles. Iteratively, we selected comments that we felt best exemplified levels of hate speech, and when we found ambiguities, used the comments to refine the definitions of each level. This allowed us to identify distinguishing features of speech and create measurement items to capture these distinctions.

Table 2.3. Example comment per theorized level of the proposed hate speech – counterspeech spectrum.

Level	Text	Hate score
5. Genocidal	Every g**k in #LosAngeles should be deported or killed.	2.69
4. Violence	#whitepower...I'll hang you n****r	2.55
3. Dehumanization	You're all a bunch of human cockroaches is what you are... You ch**ks...	1.51
2. Hostility	Learn the fucking language you fucking useless immigrant.	1.18
1. Bias	@[NAME] They are arabs. Do you need any other explanations?	0.90
0. Neutral	Go get a job at Dick's Sporting Goods and try to work at being a better person.	-0.50
-1. Supportive	I'm bi. And a good listener if you need a friend	-2.99
-2. Counterspeech	No, the chances of a muslim shooting you in America is almost nil. There are over 50K gun deaths every year christian USA...	-0.82

Example comments from each level of the reference set are shown in Table 2.3.

Labeling instrument

In order to evaluate the comments we needed one or more human reviewers (also called "annotators", "labelers", "judges", "raters", or "moderators") to examine the comment text and estimate where each comment fell on each of our theorized components of hate speech. Most labeling tasks for training data give the labeler a detailed definition of the construct and then ask them to assign a binary label to each data point (e.g. designating a block of text as hate speech or not, designating whether an image contains a stop sign). This approach has to two shortcomings: labelers cannot indicate uncertainty and if the construct has multiple components that labelers differ on, the label does not indicate which element they disagree in. The approach described below overcomes both of these issues by decomposing the construct of hate speech into multiple labeling items and by giving labelers Likert-style response options to incorporate uncertainty.

The labeling instrument we created has three types of questions: 1) identity target items, which establish whether the comment targets a protected group , 2) scale items that measure the content of the comment along several distinguishing features of hate speech, and 3) a set of demographic questions asked about the labelers. The target items and scale items

are asked for each of the comments that labelers review, and then are followed by the demographic items. One of the scale items, sentiment, is asked *before* the target items, to get one measure for items that target a non-protected group or no group at all. If no identity groups were mentioned they were not asked any remaining scale items and proceeded to the next comment. If at least one identity group was mentioned they were asked to specify the sub-identity group(s),[2] and then asked the remaining scale items. All comments were also rated on a binary hate speech item sourced from Siegel et al. (2019) to allow comparison to the current best practice in binary hate- speech measurement.

Differences in labeler knowledge and views make consistent annotations difficult to obtain. We address differences in labeler knowledge by providing a dictionary tool for niche slurs that appear in the comments we showed to labelers. Using a new dictionary, slur words were underlined in our survey user interface. If the user moved their mouse over the underlined slur word they would be shown a tooltip stating "This word may be a slur against identity group [X]" (see Figure 2.1 for an example). This user interface feature was intended to reduce response variation due to varying awareness of slur terms, as well as to make noticeable any coded slur language in the comment (for more details on the problem of covert slurs see Magu et al. 2017).

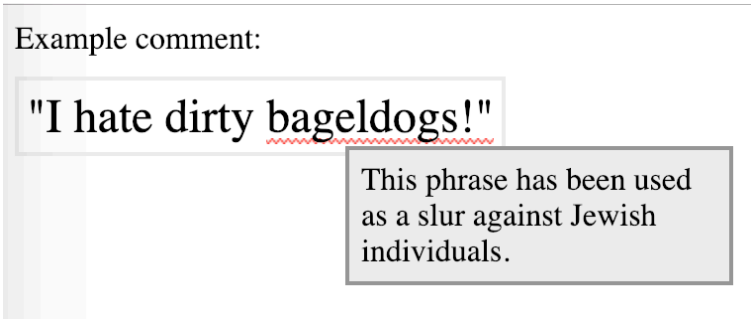


Figure 2.1. Highlighting slurs for human labelers. We highlighted known slurs for our human labelers in the annotation interface.

After rating the comments reviewers were asked a series of demographic questions about themselves, followed by an optional free response feedback item. The demographic items included the reviewer's gender, education, race, year of birth, income, religion, sexual orientation, and political ideology. We used these demographic data to investigate the associations with labeler severity, shown in Appendix 2.7.

Comment collection

We sourced our comments from three major social media platforms: YouTube, Twitter, and Reddit. We chose these platforms for their popularity, as respectively, they are used by

[2]As described in the construct theorization section, groups consisted of the categories protected under US law (e.g., religion), while the sub-identity groups are a short list of the most commonly occurring groups within these categories (e.g., Muslims). See Appendix zz for full list of identity and sub-identity groups.

73%, 22%, and 11% of U.S. adults (Perrin et al. 2019). Prior work on hate speech has often focused on a single platform, commonly Twitter, but our goal was to study hate speech in a variety of settings and to ultimately build an algorithmic model to accurately measure hate speech across multiple platforms (Fortuna et al. 2018). We used public APIs to download recent comments posted to each site. Comments were considered eligible for labeling if they were written primarily in English and were not too short (< 4 characters) or too long (> 600 characters) after removing URLs, phone numbers and contiguous whitespace. Our comment collection took place between March and August 2019.

On Reddit we collected all comments from the real time stream of the subreddit "/r/all". For Twitter, we collected tweets from Twitter's streaming API, which is a random sample of all tweets on Twitter. YouTube required additional consideration because one must first select videos and then download comments associated with the selected videos. We searched for videos within proximity of the top 300 most populated U.S. cities in order to focus on videos originating in the U.S. and most likely to contain English comments with U.S.-based authors. From those videos we then downloaded all comments and responses.

Crowdsourced labeling

Hate speech is a rare phenomenon, estimated at less than 1% of online comments when viewed as a binary outcome, so randomly sampling from the collected comments would not have been efficient. That is, the outcome in the labeled data would be highly imbalanced at $< 1\%$ hate speech and 99% non-hate speech, which would make it difficult for statistical machine learning analysis to find patterns that differentiate between hate speech and non-hate speech and costly for our labeling process. Instead, we used a sampling method that would increase the relevance of the labeled comments to our theorized levels of hate speech; we targeted an even distribution of labeled comments across our 8 levels (12.5% each). We also wanted to avoid common shortcuts to increase rates of hate speech in labeled text, such as filtering on slur terms or Twitter hashtags. Those approaches would artificially reduce the linguistic variation in the comments and allow the deep learning to learn those shortcuts (confounded associations) without capturing true patterns (i.e. causal relationships), which is known as the "Clever Hans" effect (Heinzerling 2019; Niven et al. 2019). In an effort to maximize the generalizability of our deep learning algorithm, we maintained a positive probability of selection for all sampled comments (i.e. no comments would be excluded based on their word usage).

Our sampling method relied on two dimensions for stratified sampling: 1) a relevance estimate of how likely the comment was to contain a target identity group, and 2) a hypothesis score for how hateful the comment was estimated to be. Both scores were built from a pilot set of 4,000 labeled comments, using pre-trained Universal Sentence Encoder representations (TensorFlow) plus a genetically optimized prediction head (Olson, Urbanowicz, et al. 2016). For identity prediction the genetic optimization algorithm selected a multilayer perceptron

model while for the hypothesis score it selected a random forest.[3] With each future iteration of the project we can leverage the models developed in the prior iteration to improve the stratified sampling efficiency.

We used the identity relevance and hate speech hypothesis scores to create five stratification bins: 1) irrelevant (i.e. estimated to contain no references to identity groups), 2) relevant and low on predicted hate speech score (potential counterspeech or positive identity speech), 3) relevant and moderate on predicted hate speech score (neutral), 4) relevant and high on predicted hate speech score (low or moderate intensity hate speech), and 5) relevant and very high on predicted hate speech score (violent hate speech). We heavily oversampled bins 2, 4, and 5, and undersampled bins 1 and 3. Because this stratification scheme covered all comments, each comment had a positive probability of being sampled, but we improved the likelihood of labeling comments that were some form of hate speech or counterspeech. As in a case-control study, this biased sample could be re-weighted back to the original population of comments through inverse probability weighting (Horvitz et al. 1952). We incorporated platform sample size targets such that our labeled data consisted of 40% sourced from Reddit, 40% from Twitter, and 20% from YouTube.

The sampled comments were then compiled into groups of 4 "original comments", stratified across our bins so that each group contained comments across the hypothesized hate speech spectrum. Each comment group was randomly allocated to 4 comment batches to ensure 4 ratings per comment, and each batch also included 6 reference set comments stratified across our 6 reference set levels. This design was chosen to generate a single network across all raters, which were linked through the comment groups plus the random selection from the reference set. In Figure 2.2 we show a simplified example of overlapping comment ratings that yield a single network across reviewers. This experimental design further ensured that every reviewer would receive comments across our hate speech scale; we eliminated the risk that by random chance some raters would only review comments in a narrow range of the scale.

We connected the labeling instrument, hosted on Qualtrics, to a comment batch server using web service requests in order to reserve comment batches and then mark them as completed. The comment batch server was hosted as a Python serverless function in Google Cloud with a MySQL database backend. Each reviewer was given a random comment batch of 26 comments that was not already reserved by another worker and that had not yet been completed. If a given comment batch was not completed with 10 hours it was returned to the pool of unreserved comments.

Human reviewers were recruited from Amazon Mechanical Turk to complete our labeling instrument hosted on an external site. Each labeler was given 26 comments—6 reference set comments and 20 "original comments"—to label. Median time to complete the instrument

[3]In the pilot set we used simpler scores to create the stratification bins. The maximum cosine similarity to an identity-term dictionary was used for relevance estimation in the pilot. For our pilot hypothesis score we used the Perspective API's identity attack model, which gave a predicted probability of the comment being hate speech (`https://perspectiveapi.com`).

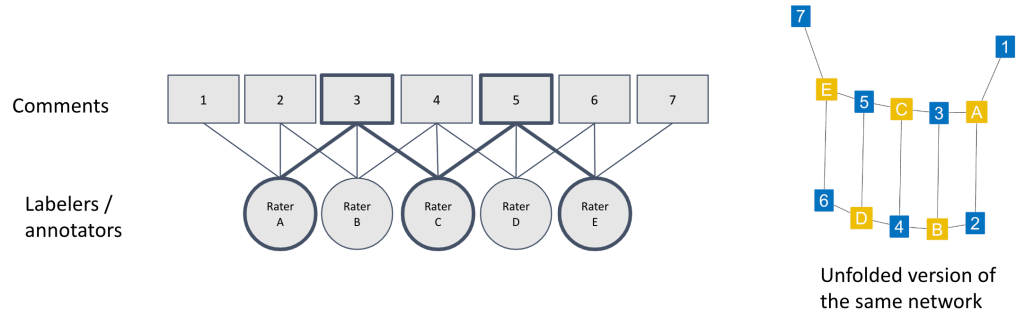


Figure 2.2. Simplified annotation example showing linkage across all reviewers.

was 49 minutes. Participants were compensated $7 for their participation in our study, yielding a median pay rate of $8.57 per hour. A manual review of the worker feedback on the task showed high satisfaction with the compensation for the task, and appreciation that the results would contribute to an understanding of social media conversations.

Constructing a scale with Rasch measurement theory

Our scaling procedure, as described in this section, converted the collection of ordinal ratings from crowdsourced human reviewers into a continuous, linear hate speech scale. We identified Rasch item response theory (IRT) as the appropriate psychometric framework to analyze the ratings and to transform them into the continuous score.

We selected the Rasch family of item response models because their theoretical properties have provably distinct advantages over other forms of IRT, leading to their elevated status as "the necessary and sufficient process for measurement" (B. D. Wright 1992). Rasch (1960, 1980) first described the requirements for objective measurement. Only Rasch models are founded on Fisherian sufficient statistics (Fisher 1934) that allow the latent variable, rater, and item parameters to be estimated separately using additive models (Andrich 2011).

Notably, Rasch models satisfy the five requirements of invariant measurement (Engelhard Jr 2013, Ch. 1) as translated into our hate speech application:

1. **Item invariance**: measurement of comments must be independent of the particular items (survey questions) used for data collection.

2. **Non-crossing comment response functions**: a more hateful comment must always have a higher probability of a more hateful response option than a less hateful comment.

3. **Comment invariance**: The calibration of items and response options must be independent of the particular comments that were labeled.

4. **Non-crossing item response functions**: A comment must have a higher (more indicative of hatefulness) expected response on an easier item (lower difficulty) compared to a harder item (higher difficulty).

5. **Construct map**: Comments, items, item responses, and raters must be simultaneously located on a single underlying continuous latent variable.

The benefit of the invariant measurement is that the resulting scale is not specific to the particular survey items, raters, or comments that we analyzed in this study. Instead, we have created a measurement system for hate speech that can be applied to future social media comments, incorporate improved survey items, and use different raters, while maintaining the ability to analyze all such objects on the original scale we have constructed in this initial work. In other words, we are able to construct a measurement device for hate speech that provides stability of meaning over a long time period, rather than an arbitrary metric that is valid only for our currently acquired data and instrument. More detailed discussions of the "Rasch rationale" are provided in M. Wilson (2004, Ch. 6) and B. D. Wright and Masters (1982, Ch. 1).

The partial credit model (Masters 1982) was the appropriate model within the Rasch family because our labeling instrument did not use the same set of response options for each item. We avoided the two-parameter or three-parameter families of IRT models such as the graded response model (Samejima 2016) or generalized partial credit model (Muraki 1992). Those non-Rasch models do not admit the invariance properties of Rasch measurement. Specifically, the estimation of item discrimination parameters (aka "slope parameter") in a 2- or 3-parameter model implicitly results in the item difficulties no longer being separable from person (comment) abilities (B. D. Wright and Masters 1982, Ch. 1, p. 8). That result can be visualized by comparing the item characteristic curves from such models and noting that they must cross, because they do not have the same shape (M. Wilson 2004, Ch. 6, p. 110 - 113).

We further avoided principal component analysis and factor analysis because they also do not transform raw scores into objective measures. Rather, they analyze the item responses as though they were already linear measures, when in fact they are only ordinal categorical ratings (B. D. Wright 1996). Moreover, neither method is sample invariant: they generate results that are intrinsically limited to the specific comments, items, and raters that were observed (Rasch 1953).

Faceted Rasch measurement

In the late 1980s Linacre extended the Rasch family of models to include judge-mediated assessments (Linacre 1987, 1989), which applies to our labeled dataset generation where a human reviewer completed a survey in which they analyzed textual comments. Treating the rater as an additional facet of the scaling procedure enabled the estimation of a rater "fixed effect" or "severity" parameter with the same hate speech scale units as the comment scores, item difficulty estimates, and item step thresholds. This severity parameter can be

viewed as an estimate of survey interpretation bias, where raters vary in how aggressively or loosely they interpret the scale items. Those rater fixed effects then no longer influence the statistical estimation of the comment abilities, item difficulties, or item steps. The result is a more objective estimate of the hate speech score for an individual comment that is independent of the severity of the raters who happen to be assigned to review that comment. Extensive details on faceted Rasch models can be found in Engelhard Jr and Wind (2018), Eckes (2015), and Linacre (1989).

With the faceted partial credit model the probability of a given response to an item can be written formally as the following equation (Linacre and B. D. Wright 2002; Eckes 2015):

$$\log \left[\frac{p_{nijk}}{p_{nijk-1}} \right] = \theta_n - \delta_i - \alpha_j - \tau_k \tag{2.1}$$

where:

- p_{nijk} is the probability of comment n being rated as response k by rater j on item i,
- p_{nijk-1} is the probability of comment n being rated as response $k-1$ by rater j on item i,
- θ_n is the ability of comment n,
- δ_i is the difficulty of item i,
- α_j is the first-order bias ("severity") of rater j,
- τ_k is the difficulty of receiving rating k relative to rating $k-1$.

Faceted Rasch models include a noteworthy implication for inter-rater (kappa) reliability: it is not essential that different raters provide the same responses to an item when analyzing a certain comment. There is a growing body of related literature showing that a single "true" labeled response is often unrealistic (Palomaki et al. 2018; Aroyo et al. 2019; Geva et al. 2019). Instead, we want within-rater consistency in item interpretation so that their estimated severity acts as a strong summary measure of their individual style of rating. In fact, for raters with very different estimated severities, we would expect them to provide different ratings on an item when analyzing the same comment - that would be consistent with the measurement model (and common sense). Reliability of ratings and unbiasedness are two distinct phenomena; for example, a given comment may exhibit high reliability due to multiple raters agreeing on an biased assessment (Henning 1996). This is a marked psychometric departure from prior studies of hate speech or other supervised natural language processing topics, which have commonly relied on inter-rater reliability as the primary quality metric for dataset labeling (Ross et al. 2016).

We used Facets software (Linacre 2019) to conduct the many-facet Rasch scaling. A sequence of four scaling estimates were conducted, in which increasing percentages of low-quality raters were removed, and response options were collapsed to reduce noise in the

estimates. The details of the rater quality analysis and response collapsing are described in the appendix.

Deep learning prediction

After scaling, we trained an algorithm to estimate a mapping from the raw text to our latent hate speech score. Deep learning has shown the best performance for this type of task, provided that sample size is not too small. The current best architectures are based on transfer learning (Pan et al. 2009) and Transformer units, such as T5, ALBERT, and RoBERTa. These architectures consist of a language model that is first trained in an unsupervised fashion on large amounts of general text, typically millions of Wikipedia articles and about 10,000 books. This teaches the algorithm the meaning of words within their context, allowing it to read new types of text.

We then conducted the supervised learning: we supplied the raw comment text as our input observations, with the ultimate goal of predicting the latent hate speech score. We made three novel changes in our supervised approach: 1) rather than predict the hate speech score directly, we instead predicted the responses to each survey item using a multitask architecture (i.e. multiple outputs within a single model) (Ruder 2017), 2) each survey item was directly analyzed as an ordinal outcome using a proportional odds model activation (Vargas et al. 2019) combined with quadratic weighted kappa loss (de la Torre et al. 2018), and 3) we supplied the rater's estimated bias (severity) as an additional non-text input to allow the model to adjust its understanding of the likely item response. The individual item response predictions could then be transformed into latent scores using the estimated IRT parameters. See Figures 2.3 and 2.5 for depictions of this architecture. We believe that this approach has never before been implemented and represents a new way of integrating deep learning with item response theory for measuring phenomena that can be decomposed into multiple items that are reviewed by human raters. We now describe the details of those steps.

The simplest supervised learning strategy would be to predict the continuous hate score directly, and to ignore the intermediate item ratings that led to that score. With this architecture the data structure would consist of one observation per comment. The estimated severity (bias) for each labeler would not need to be incorporated as an auxiliary input because the score was already debiased through the IRT scaling. The loss function could simply be mean-squared error and the architecture could consist of a Transformer-based representation subnetwork applied to the raw text, followed by one or more dense layers to learn the mapping to the continuous score. In lieu of the dense layers, the language representation could be applied to generate a summary feature vector for each observation, and traditional machine learning such as XGBoost (Chen and Guestrin 2016), BART (Chipman et al. 2010), SuperLearner ensembling (Polley et al. 2019), etc. could be applied to predict the continuous score. We implemented both the full neural architecture (dense layers) as well as the traditional machine learning structure with TPOT optimization (Olson and Moore

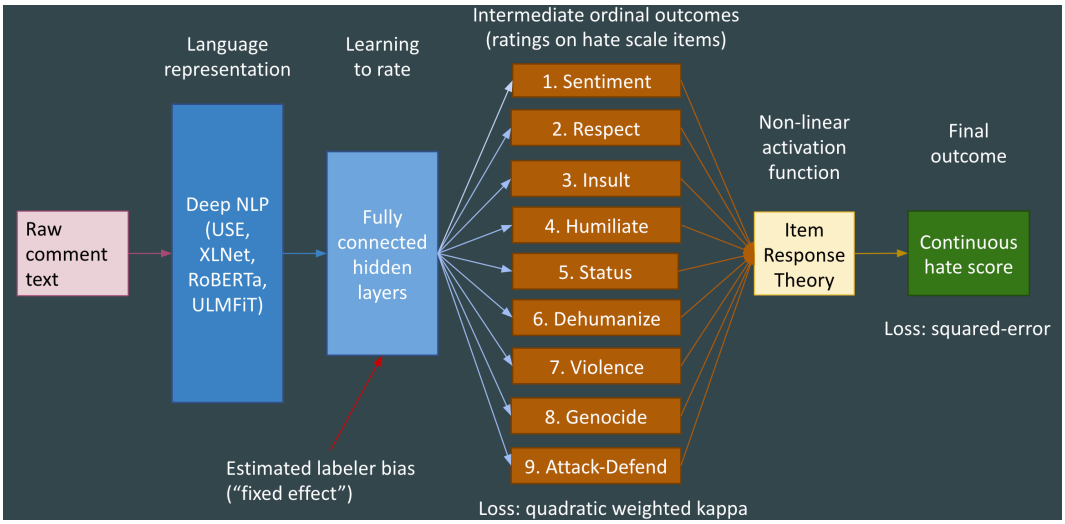


Figure 2.3. Neural architecture for predicting a continuous score with multiple intermediate outcomes, labeler bias adjustment, and IRT activation. Comment text is fed into a deep natural language processing algorithm to convert it into a fixed vector representation. Then a series of fully connected layers learns how to combine that vector representation into latent variables that can predict the response to each item on the labeling instrument. The fully connected layers also take as input the estimated rater bias for each comment rating to adjust their expected item response predictions. The predicted item responses are then transformed via IRT into the continuous hate speech score.

2019) as benchmark options.

Survey items as supervised concept learning

We hypothesized that an improved supervised learning architecture would instead attempt to predict the human rating on each of the survey items (as listed in Table 2.7). Those intermediate predictions could then be transformed into the continuous score using the IRT parameters estimated during the scaling process. The exciting insight of this approach is that the items facilitate what could be called **directly supervised concept learning**. In a typical neural architecture the final hidden layer has developed the highest level concepts that best predict the outcomes in the output layer based on the architectural hyperparameters: loss function, number of hidden units, activation function, random initialization of weights, dropout, and optimization over multiple epochs of a certain batch size. Exactly what those final concepts mean is not immediately obvious, and is the result of a greedy stochastic process that reflects a local optimum. In contrast with our method we know from our theorization and development of the construct the exact concepts that need to be learned as well as the nonlinear function that transforms those constructs to the continuous score, and we have labeled data for each of those concepts: the survey item responses from the human

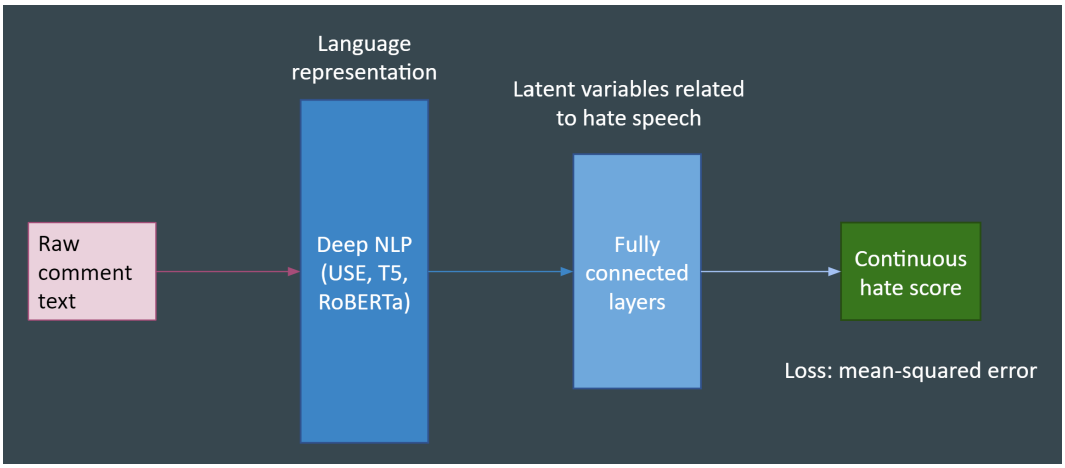


Figure 2.4. Neural architecture for predicting a continuous score directly. Comment text is fed into a deep natural language processing algorithm to convert it into a fixed vector representation. One or more hidden layers, or traditional machine learning algorithms, learn a function to map that fixed representation to minimize mean-squared error loss for predicting the continuous hate score.

labelers. This is a powerful shortcut in the supervised learning process that largely eliminates the need for architecture search or hyperparameter optimization in that final concept layer of the neural network. We know how many hidden units there are, the form of the loss function for those units, and the nonlinear activation function to transform those units into the final output. And it provides direct supervision during the optimization process for those concepts: we don't have to rely on backpropagation of errors from the final continuous output.

Multitask architecture

While there could be a separate model for each item, a multitask architecture was advantageous for three reasons: efficiency, generalizability, and convenience. Multitask architectures can improve efficiency and generalization because they bias the network to learn a shared representation of concepts that explains multiple related outputs (Caruana 1997; Goodfellow et al. 2016, §7.7). Multitask architectures are gaining increasing adoption across deep learning tasks, include face analysis (Ranjan et al. 2017), language modeling (CITE), self-driving cars (Karpathy 2019), etc. We examined the relatedness of items in our scale through a correlation heatmap (Figure 2.6) and found strong levels of correlation, supporting the likely benefit of multitask learning. Multitask items offer convenience through packaging multiple outputs into a single architecture, reducing the lines of code needed for training and prediction.

Ordinal loss function

Each item served as an output in the network, and the label was an ordinal Likert-style variable with 5 response options typically, such as: {strongly disagree, disagree, neutral, agree, strongly agree}. The loss function for each output (task) would benefit from acknowledging the ordinal nature of the outcomes: predicted probability mass assigned farther from the true label is worse than probability mass allocated to the adjacent label(s) (Hou et al. 2016). For example, when a human rater labels a comment as "strongly disagree" on the "calls for violence" item, the predicted probability of "strongly agree" should contribute much more to the loss than "disagree", because it is a much worse prediction. This has been recognized as a desire for unimodal probability distributions when conducting ordinal classification (Costa et al. 2008; Beckham et al. 2017). Cross entropy, the standard loss function used for discrete labels, does not do this. It only examines the probability placed on the true label and encourages that probability to be maximized. That is reasonable when predicting categorical variables where there is no ordering to the labels. But with ordinal variables we could potentially achieve greater sample efficiency by using a continuous form of the weighted kappa loss as proposed by (de la Torre et al. 2018). Table 2.4 displays the penalization matrix for an item with five possible responses and a quadratic distance scaling factor.

		Observed item label				
		1	2	3	4	5
Predicted	1	0	0.0625	0.25	0.5625	1
	2	0.0625	0	0.0625	0.25	0.5625
	3	0.25	0.0626	0	0.0625	0.25
	4	0.5625	0.25	0.0625	0	0.0625
	5	1	0.5625	0.25	0.0625	0

Table 2.4. Quadratic penalization matrix as part of weighted kappa loss for items with 5 responses.

The penalization matrix was constructed using the formula:

$$\omega_{i,j} = \frac{|i-j|^p}{(C-1)^p} \tag{2.2}$$

where $\omega_{i,j}$ refers for the penalty for probability mass assigned to response i when the true label was response j, p is a scaling factor for the increase in penalty as distance increases and is 2 typically (quadratic penalty), C is the number of response options for a given item and is 5 typically, and $\omega_{i,j} \in [0, 1]$ (Vargas et al. 2019).

The quadratic penalization matrix is the ω term in the **continuous weighted kappa** (CWK) loss function shown below (ibid.).

$$\text{CWK} = \frac{\text{Observed}}{\text{Expected}} = \frac{\sum_{k=1}^{N}\sum_{q=1}^{Q}\omega_{t_k,q}P(y = C_q \mid x_k)}{\sum_{i=1}^{Q}\frac{N_i}{N}\sum_{j=1}^{Q}\left(\omega_{i,j}\sum_{k=1}^{N}P(y = C_j \mid x_k)\right)} \tag{2.3}$$

The numerator iterates over each observation indexed by k and sums the predicted probability for each possible response q, weighting the loss by the corresponding entry in the ω penalization matrix given the observed label t_k. The denominator normalizes that value to [0, 2] bounds.

Ordinal activation function

We paired the ordinal-focused loss function with an intermediate item-specific output layer that generates a probability distribution intended for ordinal outcomes: the proportional odds model (McCullagh 1980; Agresti 2010). The proportional odds model is a member of the class of cumulative link models (CLMs), which model ordinal outcomes by estimating the cumulative distribution as a linear function after applying the inverse link (Vargas et al. 2019).

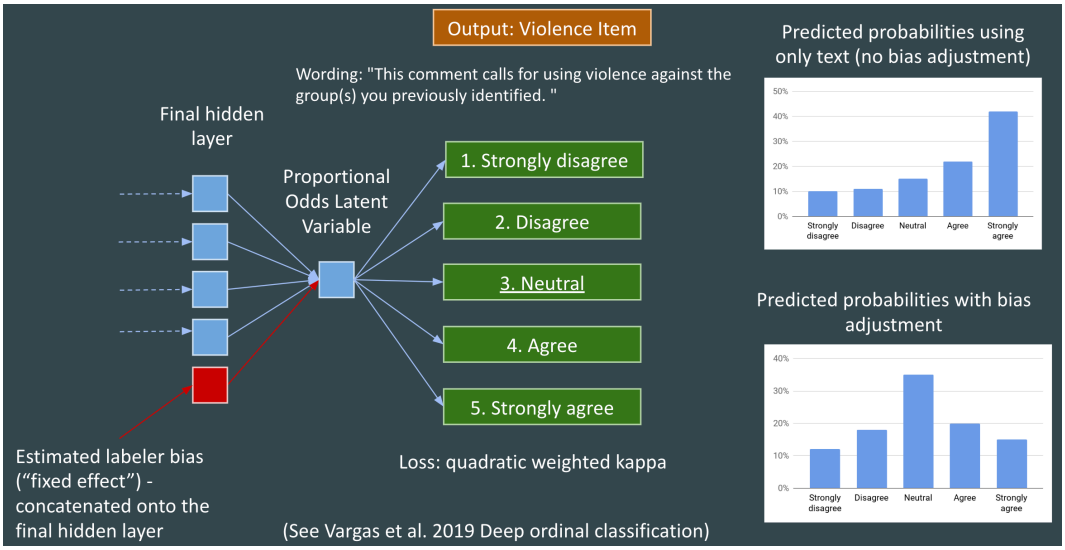


Figure 2.5. Example of ordinal classification for the violence item. Labeler bias is concatenated onto the final hidden layer as an auxiliary input. Weights connect from the final hidden layer to a proportional odds latent variable model, allowing bias to scale the prediction up or down the ordinal output labels. The proportional odds model outputs a unimodal probability distribution that respects the ordinal structure of the item labels. Loss is evaluated using the quadratic weighted kappa metric, which discourages probability mass far from the observed label.

Partial credit transformation of predictions

Once the multi-task deep learning model has been trained to generate item-level predictions (i.e. probability distribution over each response option), those predictions need to be fed into an anchored faceted Rasch model to be converted into the predicted hate score. The simplest approach would be, for each of the 10 items, to predict the response option (label) with the highest estimated probability. For example, if the model estimated that the genocide item would be answered as "yes" at 10% probability and "no" at 90% probability, the model's label prediction would be "no". So the model would generate estimated item labels for each comment, and those predicted ratings would be scaled by running the partial credit scaling procedure with item and response parameters anchored (fixed) to the values from the original faceted Rasch scaling. The measurement model for this deep learning scaling differs from the Rasch scaling we used to create the continuous outcome variable because we have only a single rater - the deep learning model itself. That is the reason that we use the partial credit model for this step, rather than a faceted Rasch model that estimates a severity parameter for each rater.

Plausible value sampling

There are two key downsides to the partial credit scaling as described:

1. The number of unique predictions generated by the partial credit model is limited to the cardinality of the raw scores. This is because the raw score (the sum of the individual item ratings) is a sufficient statistic for the continuous latent variable generated by Rasch measurement models. In our case the final item setup had 33 possible raw scores (0 - 32), so only 33 unique point estimates could be generated by the partial credit model transformation. Although certainly more granular than predicting a yes/no or ordinal label, that seems somewhat coarse when covering a continuous spectrum from -8 to +5, which could limit the predictive performance of this version of the model.

2. Possibly more importantly, in selecting the most probable rating we discard the information contained in the probability distribution for each item, i.e. how certain the model is about a predicted item rating for a given comment. That information seems quite valuable, and is missing from the original labeled data from the human raters - we only know the single response option that they selected (presumably the highest probability response option from their mental model of the rating task).[4] Incorporating that probability information will also improve the sensitivity of overall model performance to improvements in the multi-task item predictions: architecture improvements will translate into improved probability predictions more efficiently than improved label (rating) predictions.

[4]This begs the question: could we ask human raters to provide a probability distribution over the response options, rather than selecting a single rating? Perhaps this has already been proposed or studied.

How best to solve these problems then, generating more precise predictions that incorporate the confidence information from the predicted probability distributions? We chose to take a plausible value sampling approach, which resolved both issues. Rather than select the most probable response for each item, we sampled many possible item ratings for each comment based on the probability distribution for each item. Each "plausible value" for an item rating was selected at random from the possible response options, with discrete probability equal to the model's predicted probability for each response option. In other words, we asked the model to rate each comment many times, and to select each item rating by a probability-weighted sample of the response options. Those replicated ratings were then scaled simultaneously in the anchored partial credit model to yield an estimated hate score for each comment. We tested different replication counts (1, 2, 4, 8, 10, 16, 24, 32, 64, 128, 256) to examine the trade-offs of predictive performance and computation time.

2.4 Results

We now report on our observed results from the evaluation of the allocation of comments to raters (judging plan), application of faceted IRT to our labeled comments (scaling), and our accuracy at predicting the hate score using deep learning on the raw comment text.

Network analysis for comment batch annotation

We evaluated the network linkage across the raters, original comments, and reference set comments. We confirmed that our batch creation procedure generated a single linked network, with no disjoint subsets, that would allow the estimation of rater severity ("fixed effects"). The diameter of the network was 6, meaning that any two points were connected by traveling across no more than 6 edges. The average distance was 3.6, meaning that any two nodes were typically about 4 edges apart.

Rasch measurement theory scaling

Our faceted partial credit model achieved a case (comment) reliability of 0.94. Our estimated rater separation reliability was also 0.94, suggesting that our judging plan resulted in high accuracy at estimating the individual severity for the labelers. Following Linacre (1999), we confirmed that for each item the average hate score increased monotonically with more hateful response options.

The calibrated Wright Map showing our scale across comments, items, item steps, and raters is shown in Figure 2.7.

Our estimated item difficulties (Table 2.5 and Figure 2.7) were consistent with our hypothesized item difficulties, which speaks to the construct validity of our instrument. Our item fit statistics exhibited reasonable fit to our assumed Rasch model. Both infit (inlier sensitive) and outfit (outlier sensitive) mean-squares fell within the [0.7, 1.3] heuristic target

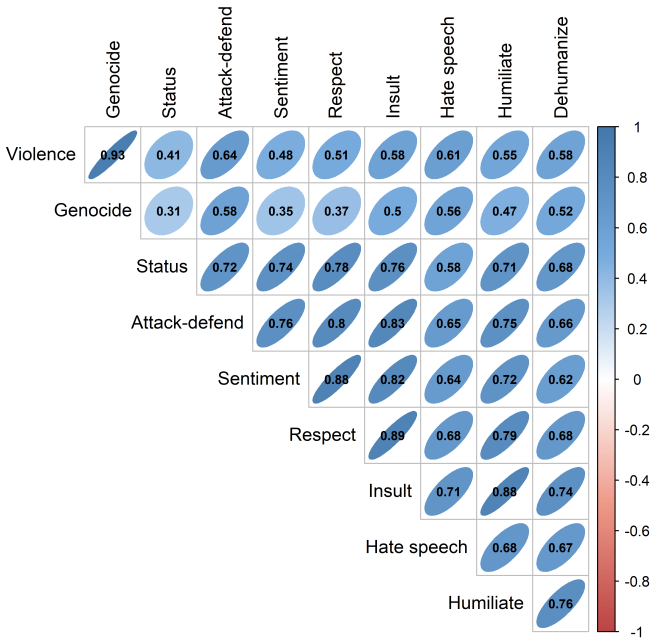


Figure 2.6. Polychoric correlations between the items in our study. All items have a positive correlation with each other, varying from 0.31 to 0.93. The most highly correlated pair is genocide and violence (0.93), followed by insult and respect (0.89).

for statistical agreement between the proposed Rasch model and the observed ratings (Eckes 2015, 5.1.2, p. 77).

We reviewed the scaling results on our reference set comments, which were each rated between 20 and 40 times. Figure 2.8 displays the ability point estimates for each reference set comment grouped into its qualitative level. The average score within each level showed the expected monotonic increase in hate speech for increasing levels. However, the level pairs 2 & 3 and 4 & 5 did not show large absolute differences in their average hate speech scores. We also noted that some comments appeared to be better moved into adjacent levels based on their hate speech scores. These results imply that we should revisit our theorization for the reference set, review the criteria for the comments to determine if they are better placed in adjacent levels, and consider substituting ambiguous comments with more clearly exemplary comments. We may want to simplify the reference set by merging levels 2 & 3 and levels 4 & 5.

We examined the distribution of the hate speech score across our three platforms (Figure 2.9). YouTube and Reddit looked very comparable. Twitter showed a similar distribution but shifted to the left, with a noticeable reduction in scores at the high end of our hate speech scale. This result suggests that we may need to improve the allocation of Twitter comments into batches in order to increase the percentage of comments on the hateful side of the spectrum. Our observed results may also indicate that Twitter already conducts some sort of automated filtering of hateful comments before allowing the comments to be down-

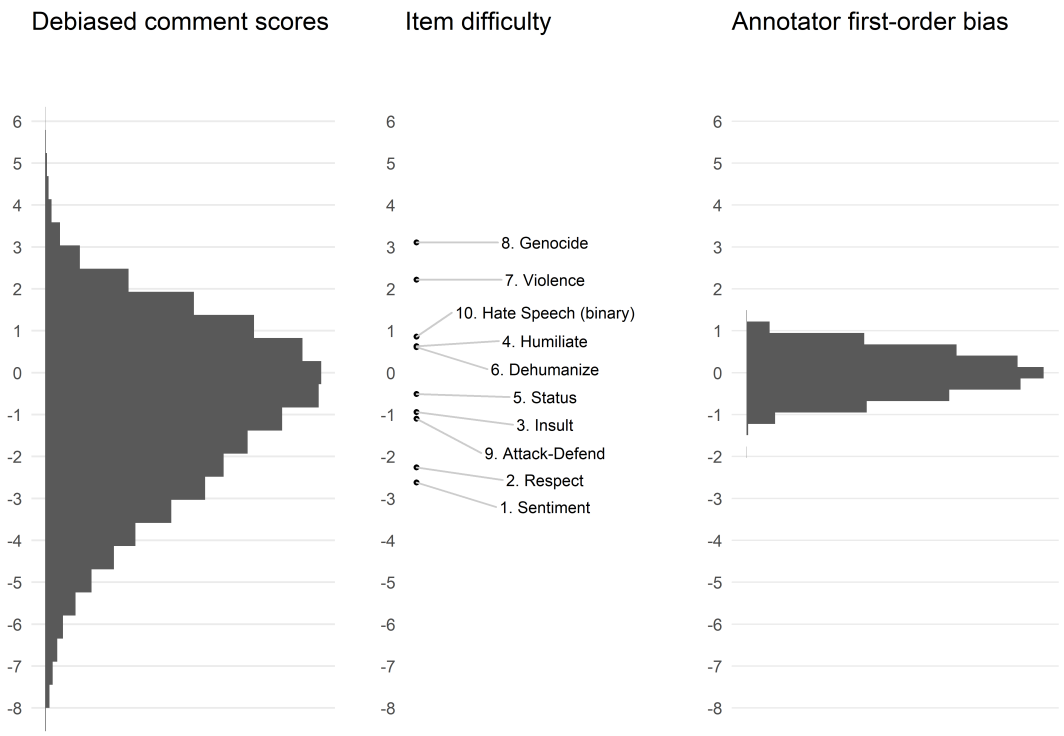


Figure 2.7. Calibrated Wright map. Comments. Histogram of the comment ability estimates from weighted likelihood estimation. **Items**. Item difficulty parameter estimates. **Raters**. Histogram of the rater severity estimates

Table 2.5. Item parameter estimates and fit statistics. MnSq = mean-squared, discrm = discrimination, ptmea = point-measure correlation.

Item	Difficulty	Infit MnSq	Outfit MnSq	Discrm	PtMea
Sentiment	-2.62	1.04	1.05	1.00	0.83
Respect	-2.26	0.94	1.01	1.09	0.85
Attack-Defend	-1.10	1.05	1.07	0.93	0.80
Insult	-0.94	0.96	1.04	1.04	0.84
Status	-0.51	1.04	1.19	0.93	0.65
Dehumanize	0.61	1.09	1.09	0.87	0.60
Humiliate	0.63	1.10	1.05	0.90	0.72
Hate speech (binary)	0.86	0.97	0.89	1.04	0.62
Violence	2.22	0.91	0.89	1.09	0.51
Genocide	3.11	0.85	0.90	1.12	0.44

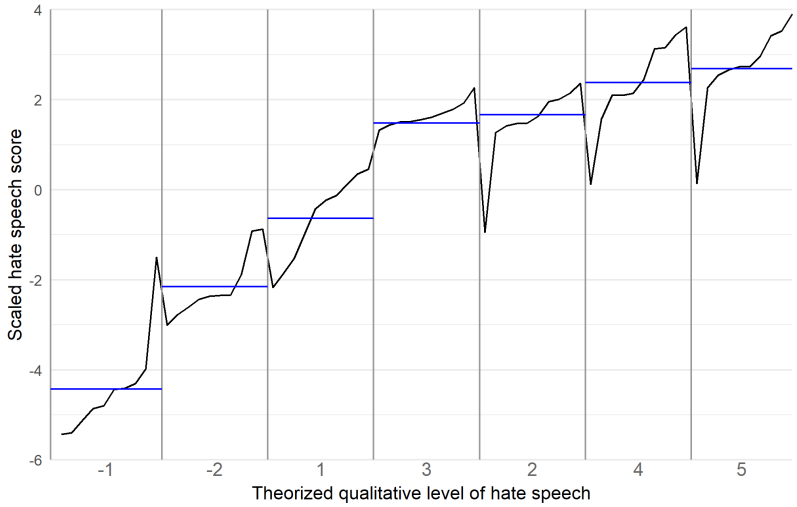


Figure 2.8. Evaluation of scaling on reference set Ability estimate for each reference set comment. The horizontal blue line is the average score within the theorized qualitative level.

loaded through their API, or that the Perspective API is less accurate on Twitter comments. Alternatively, this may suggest differential item functioning of our labeling instrument for Twitter comments, if they have linguistic characteristics that cause the raters to interpret them differently from comments sourced from YouTube or Reddit. It is important to note though that these results reflect the unweighted distribution of our labeled sample, which was intentionally skewed during the comment collection and batch creation process. Therefore these distributions should not be seen as population estimates of the true score distribution on each platform, but rather as descriptions of our training data. Future work will apply weighting corrections to the training data to estimate population-level parameters.

Deep natural language processing

After scaling, we used supervised deep learning to estimate a mapping from the raw text to the continuous hate speech score. Our current best models use a RoBERTa-Large contextual language base to process the raw text into a vector representation (Y. Liu et al. 2019), and include fine-tuning of the language representation subnetwork . Results from our model testing are shown in Table 2.6. Direct prediction of the continuous hate score has currently achieved the lowest root mean-squared error (RMSE), although our proposed multitask ordinal network that is transformed via IRT has achieved comparable performance with the benefit of explainability. In the multitask outcome representation we did not find a benefit from ordinal modeling of the items compared to categorical modeling. In supplemental results (not shown) we did find that the multitask architecture achieved lower mean absolute error compared to directly modeling the continuous hate score.

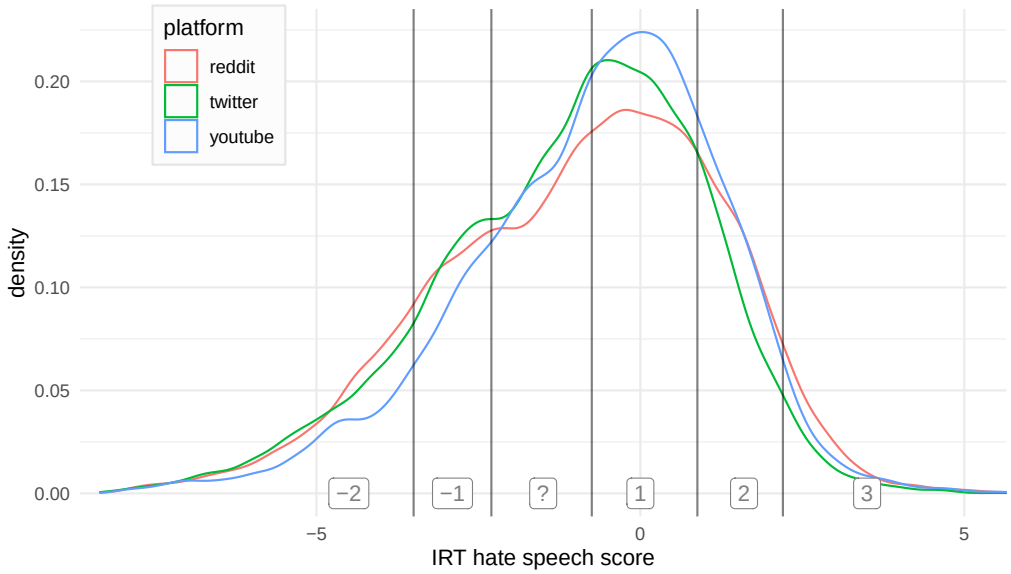


Figure 2.9. Scaling results by platform. Density of the ability estimates by platform. The lines are the proposed thresholds separating the revised theoretical construct levels.

2.5 Discussion

Heavily grounded by theoretical insights, our empirical results showed that we were successful in both major sub-projects of this work: 1) creating a 10-item labeling instrument processed with faceted Rasch modeling to yield a continuous scale for hate speech, including counterspeech, and 2) predicting that scale on unseen social media comments using Transformer-based natural language processing. That prediction could be accomplished either by directly predicting the hate score from the raw text, or by taking advantage of the Rasch transformation by first predicting the human ratings on the ten constituent components of hate speech (i.e. the multitask architecture), and then aggregating multiple possible ratings of each comment (plausible values) into a predicted hate score with a fixed-parameter Rasch scaling. The two prongs of this work allow arbitrary text to be placed on an interval spectrum ranging from genocidal hate speech on one extreme to supportive identity speech on the other extreme. In sum, we successfully created an initial measurement system for hate speech. Future work will no doubt provide further improvements to the scale creation and prediction algorithm, as well as the source data collection and the comment labeling system.

While we have focused on our hate speech problem of interest, the methodology we have developed is applicable whenever humans review data to make a judgment, and that summary judgment could be based on multiple component parts. That data could be images or video just as well as text, or other data structures such as time series. Within natural language processing, sentiment analysis is one of the most widely used supervised learning problems; we suspect that the application of our method to measure sentiment would lead

Table 2.6. Supervised learning results.. FE = feature extraction (i.e. freezing the weights in the language representation layers). FT = fine-tuning of a transformer-based architecture, as compared to feature extraction. WWM = whole-word model, referring to the form of masked language modeling. CV = cross-validation. RMSE = root mean-squared error. Corr = linear (Pearson) correlation.

Prediction Algorithm	**Outcome Representation**	**Trained Parameters**	**CV RMSE**	**CV Corr.**	**Training comments**
Outcome mean baseline	Continuous	1	2.061	0.00	42,000
Google Jigsaw Identity Attack	Binary + calibration[a]	Unknown	1.868	0.44	∼100,000
In-domain embeddings + Bi-LSTM	Continuous	9,121	1.342	0.76	37,817
Universal Sentence Encoder	Ordinal + IRT	33,539	1.286	0.78	37,817
Universal Sentence Encoder	Continuous	32,897	1.245	0.79	37,817
BERT-Large WWM FT	Continuous	335,142,913	1.142	0.82	28,571
RoBERTa-Large FT	Ordinal + IRT	355,360,769	1.126	0.84	28,571
RoBERTa-Large FT	Categorical + IRT	355,360,769	1.117	0.84	28,571
RoBERTa-Large FT	Continuous	355,360,769	1.078	0.84	28,571

[a]Because this is a probability prediction for a binary outcome, we linearly calibrate the probability output to our continuous outcome by fitting an OLS regression on the training set with the probability output as the only feature, then apply the prediction of that regression estimator to the validation data.

to substantial improvements in that subfield. Automated essay grading might be an even more straight-forward application of our method. Possible video applications include the automated evaluation of technical skill derived from videos of surgical procedures, where a multi-component labeling instrument is already used by human judges, called "OSATS" (Martin et al. 1997), and the assessment of the quality of surgical wound closure (Blencowe et al. 2019). Numerous other applications could likely be developed in future years.

Limitations and future work

Our work thus far is promising but with some known limitations. The ten items used for scaling deserve further comparison and development; we may be able to remove certain items or add additional items to increase reliability. During the scale development our estimated item fit statistics encouraged the collapsing of response options to improve invariance. This suggests that reviewers had difficulty with consistently differentiating between response options for several items. This is quite reasonable given that our response options were primarily

Likert-style "strongly agree" to "strongly disagree" - those options are inherently subjective and ambiguous. Additional theorization, qualitative review, and pilot testing of improved response options could lead to better measurement characteristics, particularly for the items with greater estimated difficulty such as genocide and violence. Clearer, more objective response options that are consistently interpreted should then result in reduced variance in the rater severity parameter, and increased precision for the latent variable. That increased precision may better separate the latent variable estimates for the different theorized levels in our reference set. Further refinement of the reference set and theorized levels may help to provide even more granular distinct levels of measurement. Incorporation of the rater bundle model into our estimation procedure will also provide a more accurate estimate of instrument reliability (M. Wilson and Hoskens 2001).

In addition to those incremental improvements, our leveraging of item response modeling opens up powerful methods for the analysis of bias and fairness. The field of measurement has long examined bias in exam questions through the lenses of differential item functioning (DIF) and differential rater severity (Myford et al. 2003). DIF methods can test if one of our 10 components of hate speech does not operate consistently for certain subsets of comments. For example, if reviewers are more likely to rate comments with a racial dialect as being disrespectful (Sap et al. 2019), we can statistically identify the issue and work to create alternative item wording that mitigates the bias. If certain individual reviewers interpret the items more harshly for comments that include sexist speech, we can also identify that statistically and correct for that bias in our analysis. These types of analyses are standard practice when analyzing item response theory scales, and can be executed within a formal statistical framework that is comprehensive for the wide variety of identity groups that are included in our labeled data. Evaluation and improvement of bias & fairness will be a significant part of our future work in this line of research.

With regard to the deep learning modeling, we have not yet reached the limits of supervised performance using our existing training data. The language representation backend can be continually upgraded as new language models are developed, such as T5 (Raffel et al. 2019). In-domain pretraining of the language model on our own corpus of social media comments would likely improve performance, which is an upgrade we are currently exploring. The multi-task architecture, and implementation of ordinal loss and activation, both likely have untapped performance improvements, such as those that might be discovered through neural architecture search or through customized task weighting (Klyuchnikov et al. 2020). The integration of the IRT transformation with the multi-task predictions bears substantial future experimentation to answer key questions. Can we approximate the IRT transformation through an integrated neural subnetwork? Is plausible value sampling the best way to capture uncertainty in the predicted ratings? What about using the item ratings as a pretraining objective for a second-stage model that predicts the continuous score directly? Robustness to misspellings or other adversarial orthography would be a helpful improvement (L. Sun et al. 2020). Lastly, numerous low-level optimization improvements are likely possible for the training of deep models, including learning rate scheduling, next-generation optimizers (e.g. Yogi), and differential learning rates.

We have a number of longer term goals for future work in our hate speech research agenda. Most urgently, there is a need to apply this measurement technology to causal inference problems such as estimating the effects of policy changes, current events, and user interface interventions on hate speech, as well as purely descriptive work to report on temporal hate speech trends. Extending our current model to English-speaking countries outside of the United States (e.g. the United Kingdom or English-speaking African states) will require a broader theorization of hate speech to alternative cultures and configurations of vulnerable populations. Expanding our platform sources will likely continue to be fruitful; we hope to include data from Facebook, Instagram, Wikipedia, Twitch, and WhatsApp in later work. Expanding to additional languages, especially Sinhala, Khmer, Arabic, Hindi, and Portuguese, will allow our work to be relevant to developing countries where ethnic conflict, genocide, and extremist violence may be more overt than in the United States. Releasing our models to select partners through an API will facilitate incorporating them into browser plugins, social media platforms, and other user interface interventions. We also plan to apply our models to extremist and radicalizing literature, such as Hitler's *Mein Kampf*, to better understand the role of hate speech in literary works.

2.6 Conclusion

In this paper we described the development of a novel, holistic methodology for measuring hate speech in a scalable, debiased, explainable manner. Based on prior literature, we theorized eight qualitative levels on a scale from genocidal hate speech to counterspeech and collected empirical observations as examples of each level (the *reference set*). We developed a labeling instrument to record ordinal ratings on 10 components of hate speech through a reviewing process. We collected online comments from three major social media platforms (YouTube, Twitter, and Reddit) and sampled them in such a way as to focus our labeling on comments more likely to be hate speech or counterspeech, but maintaining generalizability by ensuring that all collected comments had a positive probability of selection in our sampling procedure. We created a crowdsource labeling procedure to allocate comments to reviewers and yield a network linking all reviewers to each other through overlapping comment reviews, facilitating the estimation of the survey interpretation bias of each reviewer (a rater "fixed effect"). We fit the faceted Rasch partial credit model to create a sample-invariant scale for hate speech that placed comments, survey instrument items, and raters on the same metric, and adjusted the estimated comment hate speech score for the estimated survey interpretation bias of the raters who happened to rate that comment. The statistical diagnostics from the Rasch model allowed us to evaluate the quality of each labeler and remove crowdsource workers with low-quality responses. Finally, we applied supervised, multitask deep learning and IRT nonlinear activation to learn an estimator that maps raw text to the hate speech score in a robust, explainable manner. That deep learning model was encouraged to gain a more general understanding of language through training on data from three separate social media platforms.

Separately, each of these steps represents a novel contribution to the hate speech literature. In combination, we believe our methodology proposes a paradigm shift in the understanding and measurement of hate speech, and in supervised learning of human-labeled data more broadly. We hope that our work will encourage other researchers to adopt *Constructing Measures*-style theoretical development & measurement in the study of complex social phenomena, including a transition from dichotomous or ordinal outcomes to continuous, linear scales estimated via Rasch-based item response modeling, adjusted for reviewer bias, and integrated into explainable multitask deep learning architectures. For future updates on our project, data, and models, please visit hatespeech.berkeley.edu.

Funding

This work was funded by the UC Berkeley D-Lab, the Anti-Defamation League, the Berkeley Institute for Data Science, and Google Jigsaw.

Acknowledgments

We thank Nora Broege, Laura Jakli, Ben Gebre-Medhin, Christopher Hench, Simal Ozen Irmak, Aniket Kesari, Renata Barreto-Montenegro, Aaron Culich, and Kimberly Izar for their valuable contributions to the project. For research assistance we acknowledge the thoughtful work of Andy Chu, Chengzhi Huang, Jinlin He, Eshaan Pathak, and Violet Yao. For helpful comments and collaboration we are grateful to Rachel Rosen, Sebastian Raschka, Victor Vargas, Brittan Heller, D. Alex Hughes, Ahmad Sultan, Alan Hubbard, Mark van der Laan, Adam Anderson, Lucas Dixon, David Mongeau, Karen Draney, Daniel Kelley, Perman Gochyev, David Bamman, and especially Mark Wilson. We thank our item panel pilot testers and study participants for their comment reviews, particularly given that the material was often emotionally taxing.

2.7 Appendix

Scale items from labeling instrument

Table 2.7. Scale items from labeling instrument.

Item name	Prompt	Response options
Identities	Is the comment above directed at or about any individual or groups based on: (multiple selections allowed)	Race or ethnicity, religion, national origin or citizenship status, gender, sexual orientation, age, disability status, or none of the above
1. Sentiment	How would you describe the sentiment of this comment?	(5) Strongly negative - strongly positive
2. Respect	Is this comment respectful towards the group(s) you previously identified?	(5) Strongly disrespectful - strongly respectful
3. Insult	Do you agree or disagree with the following statement: This comment is insulting towards the group(s) you previously identified.	(5) Strongly disagree - strongly agree
4. Humiliate	Do you agree or disagree with the following statement: This comment is humiliating towards the group(s) you previously identified.	(5) Strongly disagree - strongly agree
5. Status	This comment states that the group(s) you previously identified is:	(5) Strongly inferior - strongly superior
6. Dehumanize	Do you agree or disagree with the following statement: This comment dehumanizes the group(s) you previously identified (e.g., by comparing them to an animal).	(5) Strongly disagree - strongly agree
7. Violence	Do you agree or disagree with the following statement: This comment calls for using violence against the group(s) you previously identified.	(5) Strongly disagree - strongly agree
8. Genocide	Do you agree or disagree with the following statement: This comment calls for the deliberate killing of a large group of people from the group(s) you previously identified.	(5) Strongly disagree - strongly agree
9. Attack-Defend	Is the comment attacking or defending the group(s) you previously identified?	(5) Strongly defending, defending, neither defending nor attacking, attacking, strongly attacking
10. Hate speech (binary)	Does this comment contain hate speech, defined as "bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics, especially when the group is unnecessarily labeled (e.g., "jew bankers", "n hooligans")." The comment contains hate speech if slurs or derogatory language are used toward any person or group regardless of whether or not the target is a member of the slur-target group. (e.g. a comment calling Trump a faggot or a n***** still contains hate speech).	Yes, no, unclear

Identity group targets from labeling instrument

Table 2.8. Identity group targets in labeling instrument. Dominant subgroups noted in italics, when present. For each list of subgroups, respondents were given an "Other" category with free response. Subgroups were randomly ordered for each respondent.

Identity Group	Identity Subgroups
Race or Ethnicity	Black or African American, Latino or non-white Hispanic, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic white
Religion	Jews, Christians, Buddhists, Hindus, Mormons, Atheists, Muslims
National Origin or Citizenship Status	A specific country, immigrant, migrant worker, undocumented person
Gender Identity	Women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)
Sexual Orientation	Bisexual, gay, lesbian, heterosexual
Age	Children (0 - 12 years old), adolescents / teenagers (13 - 17), young adults / adults (18 - 39), middle-aged (40 - 64), seniors (65 or older)
Disability Status	People with physical disabilities (e.g., use of wheelchair), people with cognitive disorders (e.g., autism) or learning disabilities (e.g., Down syndrome),people with mental health problems (e.g., depression, addiction), visually impaired people, hearing impaired people, no specific disability

Data structure for deep learning

Our data structure for the supervised deep learning treated each review of a comment as an observational unit. For a typical comment that was reviewed four times there would be four observations in our training data. Each observation would have the same raw comment text but a unique rater severity for each reviewer, plus the corresponding item ratings that the reviewer provided for that comment. This differs from prior studies which have generally aggregated the data to have one observation per comment, taking the mean or mode of the binary outcome label(s) from the reviewers, and usually discarding comments where inter-rater agreement did not exceed a certain threshold. See Table 2.9 for an example of the supervised learning data structure.

Due to this repeated measures (or hierarchical) data structure we could not use simple randomization to create training/test splits. By random chance it would be likely that some of the reviews for each individual comment would be in the training set and some in the test set, making the test set performance no longer an unbiased estimator of performance on

Table 2.9. Data structure for supervised multitask learning.

Comment Id	Raw text	Rater	Rater severity	Item ratings	Hate speech score	Training or test set
1	Example A	1	0.9	1, 3, 0, ...	1.92	Train
1	Example A	2	-0.1	0, 2, 3, ...	1.92	Train
1	Example A	3	-0.5	1, 1, 2, ...	1.92	Train
1	Example A	4	1.2	1, 4, 1, ...	1.92	Train
2	Example B	5	0.3	0, 1, 4, ...	-0.43	Test
2	Example B	6	0.4	0, 3, 4, ...	-0.43	Test
2	Example B	7	-0.6	0, 2, 1, ...	-0.43	Test
2	Example B	8	-1.5	1, 0, 2, ...	-0.43	Test

future unseen comments (i.e. due to "data leakage" across the training/test partitioning). Therefore it was important to conduct clustered randomization at the comment level when creating training/test splits, taking into account the hierarchical data structure so that reviews of a given comment would all be assigned either to training or to test, and not both.

Rater quality analysis

Crowdsourced labeling is thought to generally provide research-quality data with improved diversity over convenience samples such as undergraduate students (Berinsky et al. 2012). However, it remained important to evaluate the quality of raters that created our labeled data. A subset of workers seek to maximize their effective hourly compensation by completing Amazon Mechanical Turk tasks as quickly as possible, even though their answers may be less accurate than a slower completion speed. That is a rational pursuit of self-interest, provided that it does not lead to their work being rejected upon review. In the case of our study, and academic research in general, ethical considerations precluded us from rejecting the compensation of any study participants, even if their answers were not usable. Experienced crowdsourced workers might be aware of that protocol for academic studies, which could incentivize response satisficing. The main effects of poor quality labeler data include increased variability in the item fit statistics and ability scores, and reduced reliability estimates for the scores, raters, and items.

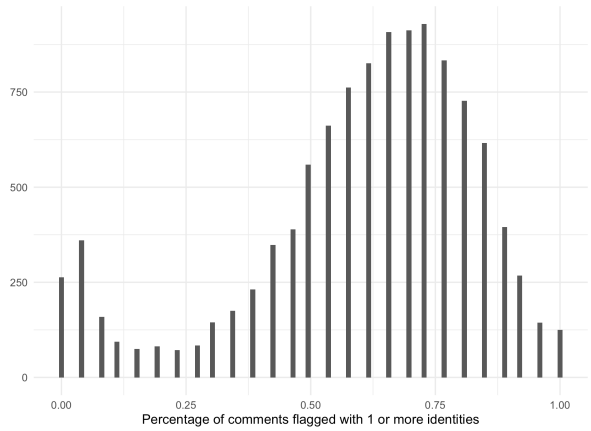
In order to evaluate the quality of raters' responses, we first examined the percentage of comments for which the rater had flagged one or more identity groups as being the targeted of the message. Our experimental design ensured that all comment batches contained six comments from our reference set, excluding the neutral level. That meant that at least 6 out of the 26 comments in the batch were known to include an identity group target (23%). The remaining 20 "original" comments were sourced from a pool of comments that were stratified on our identity model. 90% of those comments exceeded a threshold of 82% probability of containing an identity group, with the remaining 10% of comments being downsampled

from those that scored less than 82% probability. The threshold of 82% was chosen based on manual inspection. The average predicted identity probability in comments exceeding the threshold was 90.4%, and whereas for comments below the threshold the average was 27.1%. Therefore the expected identity rate in the 20 original comments was 60%, or 12 comments. The overall 26-comment batch was expected to contain roughly 18 comments with identity group targets (71%).

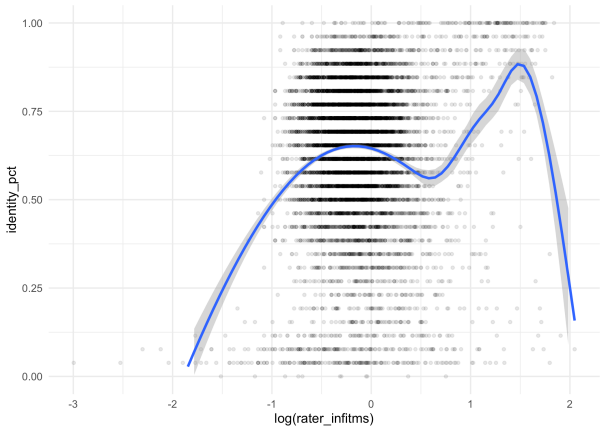
Figure 2.10a displays a histogram of the identity rate across all X,000 raters. It is a bimodal distribution, with a peak at 73%, corresponding to 19 comments out of 26 being flagged as having an identity, and at 4%, corresponding to 1 out of 26 comments being flagged as having an identity. As noted previously, each batch consisted of at least 23% of comments containing an identity group target.

We layered on a second dimension of rater quality: the *infit mean-squared* statistic, a rater fit diagnostic that is calculated during the Rasch scaling (Linacre 2002). Infit mean-squared has a minimum value of 0 and a maximum of infinity, with an expected value of 1 (or 0 on the log scale). Raters with an infit mean-squared greater than 1 had more randomness or noise in their responses than expected by the Rasch model. Those with a statistic less than 1 had less randomness than expected, suggesting that they may have favored certain response options. Values greater than 2 have been interpreted as degrading the measurement system (ibid.).

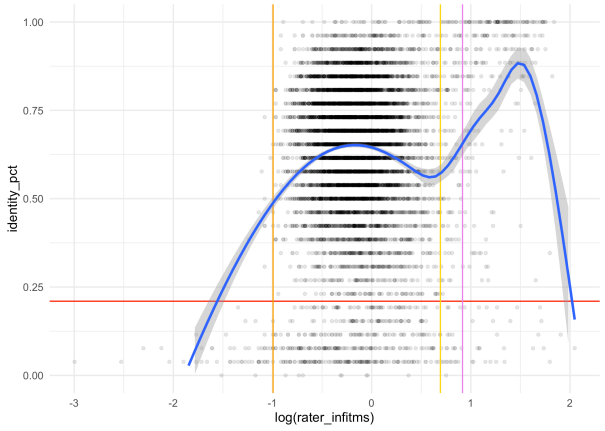
We reviewed potential exclusions on both infit mean-squared statistic and the identity percentage statistic. We chose to exclude raters with an infit mean-squared statistic exceeding 1.9 or less than 0.37, or with an identity rate less than 20%. Those thresholds led to 24% of raters being removed, leaving 8,472 as providing acceptable data quality. The post-exclusion scatter plot appeared to have a reasonable bivariate normal distribution and the smoothed identity rate became nearly flat across the infit mean-squared statistic. After those rater exclusions we re-fit the item response model and used those estimates as our primary scaling result.



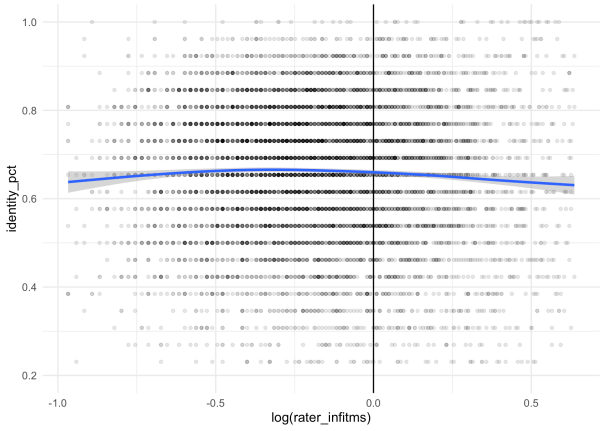
(a) Rate of flagging one or more identities in labeled comments



(b) Comparison of identity rate versus infit mean-square statistic, with lowess smooth in blue.



(c) Potential cutpoints for excluding low-quality raters



(d) Updated rater distribution after applying exclusions

Figure 2.10. Analysis of rater quality and exclusion of low-quality raters

Inadequacy of a binary hate speech item

Our inclusion of a binary hate speech item in our labeling instrument allows the comparison of our interval measure to what would be possible with only a binary item response. An initial question might be: can rater agreement on the hate speech item approximate the magnitude provided by the continuous scale? Figure 2.11 shows that rater agreement on a single hate speech item is unfortunately a poor approximation of an interval measure. The statistical association is moderate (correlation = 60%), though highly significant ($p < 0.00001$), with rater agreement explaining 36% of the variance of the measure. In addition to other benefits (invariance, debiasing, explainability, interval measurement), the continuous measure contains approximately three times the information of the single-item rater agreement approximation.

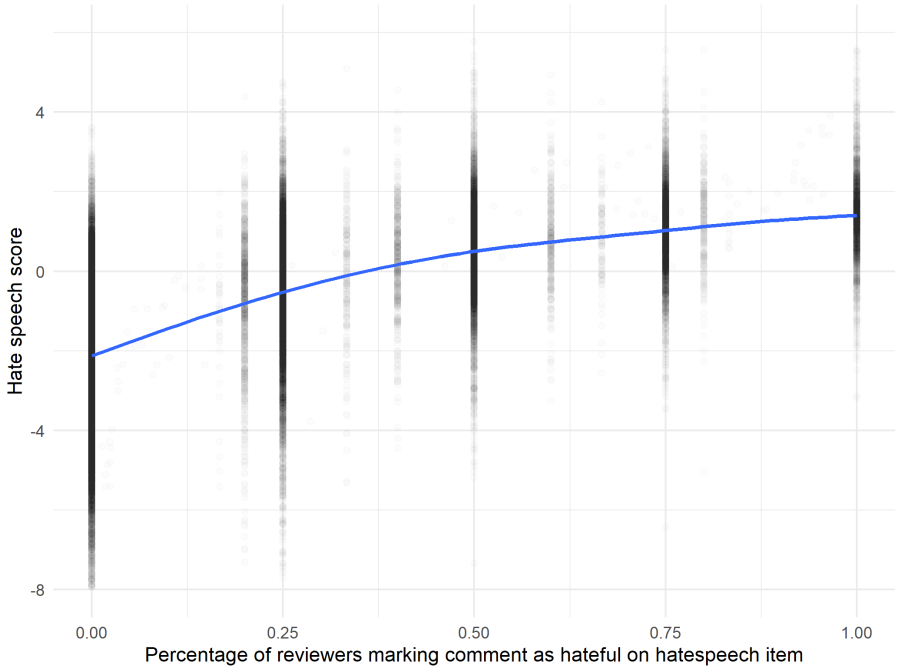


Figure 2.11. Insufficiency of a binary hate speech item. Rater agreement on a binary hate speech item fails to capture the magnitude or extremity of speech on a continuous hate speech spectrum.

Rater bias exploration

The bias or severity of the raters warranted additional exploration. What leads some raters to easily see the components of hate speech, and others to not do so? We examined the variation in rater severity by reviewing which demographic features of the raters were most associated with their estimated severity. We used ordinary least squares for a parametric

approach, random forest variable importance for non-parametric (Breiman 2001), and accumulated local effect plots for a nonparametric visualization method (Molnar 2018, Ch. 5). We hypothesized that the demographic features of vulnerable populations (i.e. people of color, women, gender non-conforming, or queer) would show negative associations with rater severity.

Our ordinary least squares analysis found that education had a significant and negative association with severity. For each unit increase on our education item a rater's severity decreased by about 0.03 points. The time it took to complete our labeling instrument was also negatively associated with rater severity, although only marginally statistically significant. This may suggest that raters more open to hate speech were more thoughtful about their answers, or were slower readers. It was noteworthy that self-reported gender, income, race, and sexual orientation variables did not show significant associations with rater severity.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.17440    0.16372   1.065   0.2878
demo_gender       -0.02739    0.04582  -0.598   0.5506
demo_educ         -0.02962    0.01331  -2.226   0.0269 *
demo_income        0.03569    0.03048   1.171   0.2428
race_white         0.01726    0.05935   0.291   0.7715
sexorient_hetero   0.07662    0.06707   1.143   0.2544
log_duration_mins -0.04129    0.02381  -1.734   0.0842 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3449 on 238 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.04333,    Adjusted R-squared:  0.01921
F-statistic: 1.796 on 6 and 238 DF,  p-value: 0.1005
```

Figure 2.12. Ordinary least squares. Education exhibits a significant negative association with rater severity, and duration of survey completion a marginally significant negative association. Overall explanatory power is low based on the adjusted R-squared and (especially) F-statistic.

For a nonparametric comparison we fit a random forest to predict rater severity by the same demographic features we examined in our OLS analysis. The variable importance results are shown in Figure 2.13. The results provide confirmatory evidence of the importance of education in explaining rater severity. Gender shows a stronger effect that was not seen in the OLS results. We did not find that duration of survey completion was strongly associated with rater severity.

Finally, we plotted the nonparametric marginal relationship of a subset of covariates to rater severity as estimated by the random forest model. The accumulated local effect

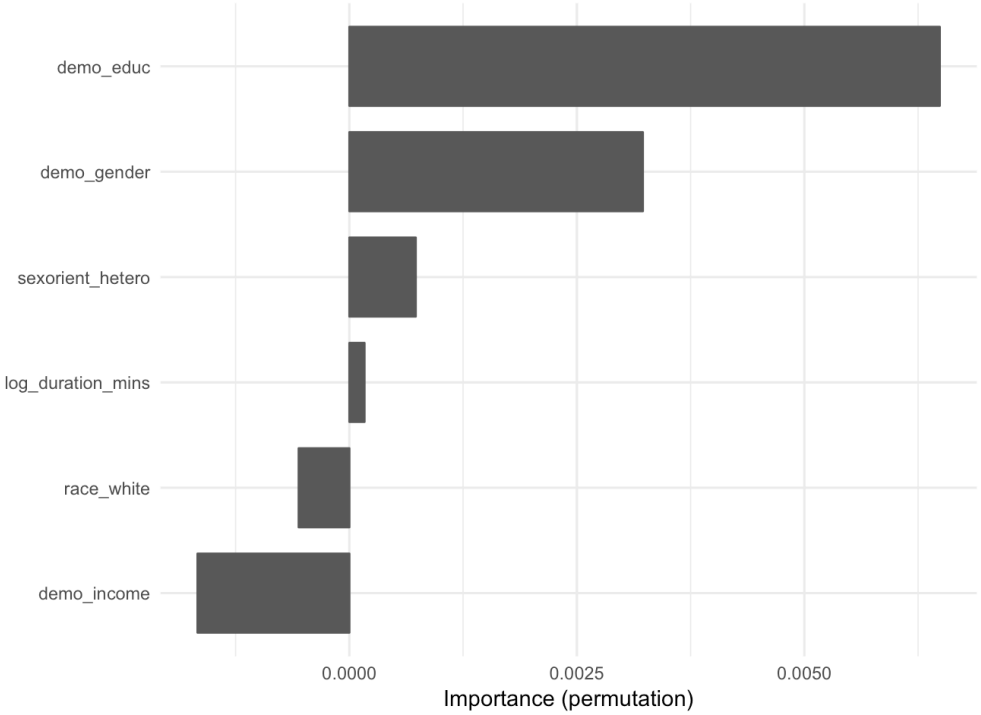


Figure 2.13. Random Forest variable importance in severity prediction. Higher values indicate stronger reliance of the covariate on the random forest's predictive accuracy. Negative values, or values close to zero, indicate that the covariate was not helpful for predicting rater severity.

(ALE) plots are a tool from interpretable machine learning that visualizes how the outcome prediction changes as a single variable varies over its range. Unlike other methods, ALE plots use only observations local to the current value of a given variable for smoothing purposes, which ensures that counterfactual covariate assignments do not result in covariate combinations that are never observed.

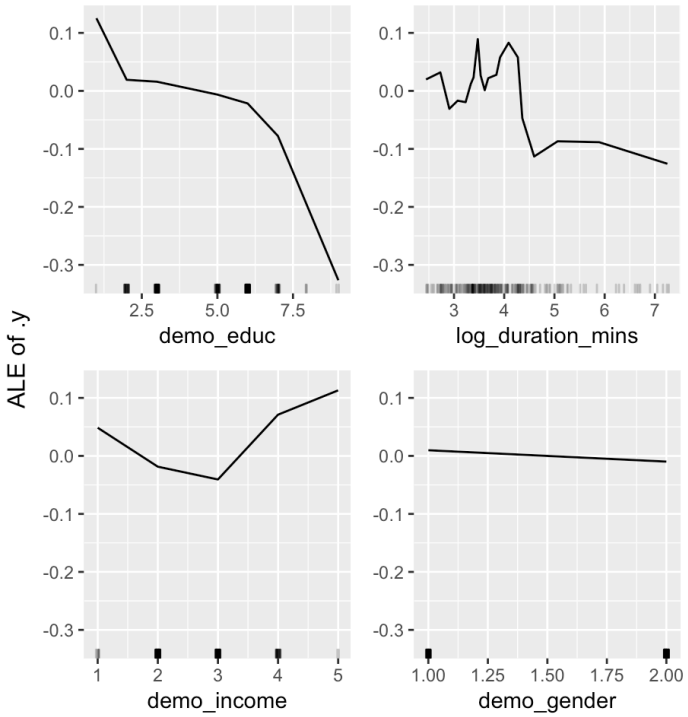


Figure 2.14. Accumulated local effects analysis

Overall we were not able to explain a large proportion of the variance in rater severity. In the future we hope to expand our collection of rater demographics to include attitudinal items, such as political ideology and/or political partisanship.

Chapter 3

Data-adaptive estimation of treatment mixtures

3.1 Introduction

Scientists typically seek to understand the effect of toxic chemicals by examining changes in one chemical exposure at a time, such as via the estimated beta coefficients of a logistic regression. The independent effect approach may poorly estimate the toxicity of chemicals that act in combination to influence disease, whether through synergistic or antagonist relationships (Braun et al. 2016). There is growing recognition that interactions between exposures are important to characterize health impacts (Binderup et al. 2003), including in areas such as air pollution (Davalos et al. 2017), endocrine disrupting-chemicals (Kortenkamp 2007; Lazarevic, Barnett, et al. 2019), childhood leukemia (Metayer, Dahl, et al. 2016), pediatrics (Henn et al. 2014), tobacco, and wildfire smoke , etc. Further indicating the importance of the topic, the National Institute for Environmental Health Sciences (NIEHS) included the study of combined exposures (mixtures) as a key goal in its 2018-2023 strategic plan[1], a continuation from its 2012-2017 strategic plan (Carlin et al. 2013). While joint effects could be estimated through multidimensional grids, that becomes infeasible as the number of exposure variables grows due to the curse of dimensionality. It also does not provide any summary measure or function that characterizes the joint relationship. As a result, recent years have seen a shift to examining joint relationships of chemical exposures as summarized through mixture modeling methods such a lasso, principal components analysis, weighted quantile sum regression (Carrico et al. 2015; Renzetti et al. 2019) Bayesian kernel machine regression (Bobb, Valeri, et al. 2014; Bobb, Henn, et al. 2018), and Bayesian generalized additive models (Scheipl et al. 2013).

In this chapter we will describe a novel methodology for analyzing exposure mixtures, or vectors of treatment variables, as data-adaptive target parameters within the targeted causal inference framework.

[1] `https://www.niehs.nih.gov/about/strategicplan/index.cfm`

Data structure

We call our observed data O with length of n, so each individual observation is O_i for $i \in \{1, \ldots, n\}$. For each unit O_i we observe a vector of adjustment variables W_i, a vector of continuous chemical exposure concentrations A_i, and an outcome variable Y_i. Y is often binary, such as case-control status, but may be continuous. We assume that O_i is independently and identically distributed. The time-ordering is assumed to be W $\rightarrow$ A $\rightarrow$ Y. W is presumed to include important observed confounders of the $A \rightarrow Y$ relationship, i.e. variables that are causally related to A and to Y, which will reduce bias in causal effect estimation by blocking backdoor paths in a nonparametric directed acyclic graph (DAG) (Greenland et al. 1999; Pearl et al. 2018). W may also contain non-confounders that are related to Y and will improve statistical power. We prefer that W does not contain strong instruments, i.e. variables causally related to A but not directly to Y, which will increase the variance of effect estimation without reducing bias. W and A should also not contain colliders, which will open back-door paths to Y and bias the effect estimates (Weisskopf et al. 2018).

Scientific goals

Our goal is to understand and summarize the potentially complex joint relationship of the exposures A on the outcome Y, adjusting for variation in W. That understanding could be applied to answer certain key questions about the exposures, such as those asked in the National Institute for Environmental Health Sciences's (NIEHS) 2015 mixtures workshop (Taylor et al. 2016). Which exposures are causally related to the outcome, and are they positive or negative contributors? Are some exposures not causally related to the outcome? What is the relative impact or toxicity of each exposure? Are there interactions between exposures in terms of antagonistic, synergistic, or other complex relationships? Can we summarize the combination of exposures into a single mixture function? If so, what are the associated risks at different levels or quantiles of the mixture? It may not be the case that a single method can answer each of these questions - it may be optimal to use different methods for the different questions. In this paper we focus initially on the last two questions: how can we create a summary mixture function, and what is the risk at different quantiles (e.g. low, medium, or high exposure) of that mixture? Depending on the estimator used for the mixture function, we may also be able to interpret the marginal effect of individual exposures.

How should we proceed? We might consider pursuing a grid strategy, in which we calculate the treatment-specific means for exposure combinations across a range of quantile values. This is similar to a factorial experimental design if we had the ability to randomize exposures. For example, we could discretize each exposure based on quartiles, then estimate adjusted treatment-specific means (i.e. adjusted absolute risks) for all discretized exposure combinations. We might then examine patterns in the treatment-specific means by the individual exposures, such as by projecting onto a linear marginal structural model.

The downside of this strategy is that it loses statistical power by partitioning the observed data into increasingly smaller subsets as the size of the exposure vector increases, or as the discretization becomes more granular. This has been called the *curse of dimensionality*. For example, if we have 4 exposure variables and convert them into quartiles, each treatment-specific mean relies on only $\frac{1}{16}$ of the observed data for the propensity score adjustment. The implication is that we will have large standard errors and wide confidence intervals when comparing the point estimates across the different combinations of exposures. Accounting for multiple testing or directly calculating simultaneous confidence intervals will exacerbate the loss of statistical power. We also have the issue of non-uniform variable importance: some exposures may be highly related to the outcome and others may have negligible or no causal relationship to the outcome. We would prefer our grid search to have increased resolution for specific exposures as they are more related to the outcome, and reduced resolution for exposures the more that they unrelated to the outcome. This will maximize power and reduce the penalty from multiple testing.

To avoid these problems, we might aggregate or project the vector of A exposures into a combined exposure that summarizes the joint (causal) effects of the A exposures in a single dimension.[2] In other words, we want to reduce the dimensionality of the A vector and minimize the loss of important information. And we want to do so in a supervised fashion such that we focus the retained information on the joint exposure subspace with the greatest counterfactual change in the outcome distribution. An unsupervised reduction, such as principal components analysis or generalized low-rank models, does not focus the reduction on our desired scientific goals. We need the exposure dimensionality reduction to be supervised so that the exposure mixture is maximally associated with the outcome variable. And because we seek to summarize the causal effects of the exposure vector, we strongly prefer that any mixture estimation be adjusted for confounders; an unadjusted mixture estimator does not seem helpful as it may simply reflect the association of the exposures with confounder variables (age, income, health behaviors, etc.). Ideally we would also allow the confounder adjustment to be nonparametric, so that we minimize residual confounding (Becher 1992). We also desire that the functional form of the exposures allows interactions between exposures, if not a fully nonparametric functional form.

Analyzing this aggregate exposure or mixture might also allow the detection of synergistic or antagonistic interactions (effect modification) among exposures, and would reduce the burden of multiple testing correction.[3]

Once we have estimated that aggregate mixture function and applied it to each observation, we could discretize the estimated mixture values into quartiles (or other quantiles) and estimate the adjusted treatment-specific mean for each quartile .A natural variable importance parameter for ranking different sets of exposure variables could then be obtained by examining the high-risk and low-risk quantiles and calculating the risk difference or risk

[2]In the future we might consider multiple dimensions.

[3]However, an effect modification framework does not quite fit here because we don't have a single primary exposure that occurs after the effect modifiers. Rather we have a set of exposures considered to have occurred approximately simultaneously, due to lack of other knowledge about time-ordering.

ratio. Such a parameter would be similar to the *varimpact* method but would not require that each quantile be examined, because the first and last quantiles should in theory always be the relevant quantiles (Hubbard and M. J. van der Laan 2016; Hubbard, C. J. Kennedy, et al. 2018). Another option is to conduct a test of the joint hypothesis that each treatment-specific mean is equal for a given mixture, or a test of the null that the continuous mixture is not related to the outcome Westling 2020. Or we could consider a grid of stochastic interventions summarized by a marginal structural model (Muñoz et al. 2012).

If we do proceed with the mixture estimation, this would be considered a *data-adaptive target parameter* (Hubbard, Kherad-Pajouh, et al. 2016; Hubbard and M. J. van der Laan 2016; Hubbard, C. J. Kennedy, et al. 2018). The term "data-adaptive" acknowledges that we do not know in advance how we should combine the exposure vector into a summary mixture, so we have to estimate that function from the data. That would then admit an efficient CV-TMLE procedure (Zheng et al. 2010), in which we leverage cross-validation to estimate the mixture exposure on a training set, apply the training sample mixture function to the remaining held-out data (test set), estimate the treatment-specific means on the test sets, and rotate that partitioning such that every observation is used separately for training and test estimation. The bias-reducing fluctuation step of CV-TMLE would also pool data across all test sets, maximizing power.

3.2 Methods

The primary three steps in our method are 1) estimate the mixture function on training data, which maps the original exposures into an aggregate, continuous-valued mixture exposure having maximum confounder-adjusted association with the outcome, 2) use held-out data (averaged across cross-validation) to evaluate the adjusted outcome mean (e.g. risk of cancer) for quantiles of the distribution of that mixture, and 3) interpret the estimated mixture function to identify the direction & magnitude of each exposure's influence, or lack of influence, as well as potential interaction between exposures. This process can optionally be run separately for groups of exposures, allowing them to be ranked based on the impact of their associated mixture.

Mixture estimation

Within the CV-TMLE framework there is wide flexibility for different methods of estimating the mixture. We first established three mixture estimation goals as guidance:

1. Causal defensibility through confounder adjustment, preferably nonparametric,
2. Incorporation of potential interactions between exposures (e.g. antagonisms or synergisms, possibly non-linear) or other complex functional characteristics of the mixture such as saturation or threshold effects, and

3. Favorable finite sample performance that maximizes statistical power.

We base our definition of the mixture score as a nonparametric generalization of one proposed earlier (Keil et al. 2020), which defined the mixture as a weighted sum of the discretized constituent chemicals that comprise the mixture. Because the assumes that the impact on the health outcome is based upon a linear sum, it is overly restrictive. We define our definition of a mixture, A, as a W-specific (A can "interact" with covariates as well has having interactions among the chemicals in the mixture and the difference the mixture makes to the estimated conditional mean beyond the covariates, W. Specifically, we define the mixture an oracle score S_0:

$$S_0(A, W) = \mathbb{E}[Y \mid A, W] - \mathbb{E}[Y \mid W] \tag{3.1}$$

The first term is a standard outcome regression that estimates the conditional expectation of the outcome given all exposures and covariates. The second term centers that estimate around the expected outcome value for each observation based only on its covariates, and not incorporating the exposure variables. We call this second term the *exposure-naive outcome regression*.

Counterfactually setting $S_0(A, W) = s$ (e.g. s is an arbitrary interval $(a, a + \epsilon)$ where $\epsilon > 0$) corresponds with drawing A from a stochastic intervention defined as the conditional distribution of A, given W and $S_0(A, W) = s$. For example, Y_s, the counterfactual outcome under setting $S_0 = s$, equals Y_{g^*} where the stochastic intervention $g^*(A \mid W)$ on A equals the true conditional distribution of A, given W and $S_0(A, W) = s$. In this paper we focus on a static intervention but in future work we plan to explore alternative causal effects, such as shift interventions.

The mixture (score) estimation procedure is operationalized as follows:

1. On the full dataset (outside of the CV-TMLE loop, since we aren't examining Y):
   - Estimate exposure regressions $\mathbb{E}[A_j \mid W]$ for all exposure variables A_j (also known as a generalized propensity score (Hirano et al. 2004))
     - Use SuperLearner ensemble estimators, or OLS for proof of concept.
   - Orthogonalize each exposure from the adjustment covariates: $A'_j = A_j - \mathbb{E}[A_j \mid W]$
2. Now within a cross-validation setup (e.g. 5- or 10-fold):
   a) Estimate an exposure-naive outcome regression $\mathbb{E}[Y \mid W]$.
      - Use a SuperLearner ensemble estimator, or OLS/GLM.
   b) Create a "meta-level" dataset comprised of stacked test sets.
      - Columns would be the outcome Y, offset $\mathbb{E}[Y \mid W]$ predicted at W_i, orthogonalized exposures A', and adjustment covariates W.

- Where the offset is the predicted value from the exposure-naive outcome regression in which the given observation was in the test set.

3. Using the meta-level dataset, run regression $\mathbb{E}\big[Y \mid A', W, \mathbb{E}[Y \mid W]\big]$ with test-set-derived offset $\mathbb{E}[Y \mid W]$.
    - This would again be a SuperLearner estimator.
    - The ideal covariates would be interaction terms of the form (and with higher orders possible): $A'_j f(W) = (A_j - \mathbb{E}[A_j \mid W]) f(W)$
    - Algorithms that don't support an offset term (e.g. random forest) can simply use $\mathbb{E}[Y \mid W]$ as a covariate. If we can force that term into the model even better.
    - Any main terms should have conditional mean zero given W, so estimators that incorporate main terms of A or W are either picking up noise (overfitting) or capturing residual information that was missed during the earlier nuisance regressions (i.e. they were underfitting).

A formal argument for the double robustness for this procedure is presented in the appendix.

The SuperLearner ensemble libraries could incorporate flexible, powerful estimators such as the highly adaptive lasso (Benkeser and M. van der Laan 2016; M. J. van der Laan and Benkeser 2018), Bayesian additive regression trees (Chipman et al. 2010), xgboost (Chen, T. He, et al. 2015), Bayesian kernel machines (Bobb, Valeri, et al. 2014), generalized additive models (Wood 2017), and Bayesian generalized additive models (Scheipl et al. 2013), in addition to lower variance estimators such as OLS, elastic net (Zou et al. 2005), and hierarchical lasso (Bien et al. 2013). The libraries could test the incorporation of preprocessing steps such as feature selection or dimensionality reduction (Z. Sun et al. 2013) using arbitrary algorithms.

The key constraint for binary outcomes is that algorithms need to support an offset set as part of the orthogonalization process. Many of the noted algorithms explicitly support an offset by default, including HAL, BART, XGBoost, GAMs, OLS, and elastic net. Algorithms that do no explicitly support an offset can be provided with the offset as a covariate, which will have a similar effect although with less power.

Effect estimation with CV-TMLE

After estimating the mixture we then apply the function to the training data to predict our estimated mixture exposure for each observation. That continuous exposure is then discretized into ordered exposure quantile intervals (e.g. 4). We then estimate propensity score functions for each mixture quantile, and a pooled outcome regression. On the test data we apply the mixture estimator and quantile discretization from the training data to generate the ordered exposure mixture, then combine with the outcome regression and propensity score to estimate treatment-specific means (TSM) for each quantile. Through

the CV-TMLE process we can pool the debiasing TSM fluctuation step by combining data from each CV-TMLE fold. The end result is a set of CV-TMLE-generated treatment-specific means for quantiles of an exposure mixture estimated on the training data.

In theory we should not need to estimate the *varimpact*-style "maximal contrast" effect parameter in which the high and low levels are estimated from the training data, because our understanding of the mixture is already ordered based on risk.[4] So this method can be seen as a general "variable set importance" method.

As an additional counterfactual estimate, we may be interested in estimating the causal dose-response curve for the mixture (E. H. Kennedy, Ma, et al. 2017; M. J. van der Laan, Bibaut, et al. 2018; Westling 2020).

Interpretation

Our initial focus is the estimation of the mixture itself and then estimation of risk at quantiles of that mixture. Yet interrogating the mixture function to understand how individual exposures, or subsets of exposures, influence the estimated mixture function can be of substantial scientific interest. That interrogation could entail five aspects of influence, in descending order of priority:

1. **Relevance**: Which exposures contribute to the mixture and which do not? (Westling 2020)
2. **Directionality**: For relevant exposures, is the exposure positively or negatively related to risk, or both?
3. **Relative magnitude**: What is the relative magnitude (potency) of impact of each exposure?
4. **Interactivity**: For each exposure, is the exposure's influence on the mixture moderated by the magnitude of other exposures? Can we reject the null hypothesis of additivity?
5. **Dose-Response**: What is the approximate shape of the dose-response curve for each exposure? Can we reject the null hypothesis of linearity? (ibid.)

Answers to these questions will improve our understanding of the etiology of the causal system, and will allow comparison to prior theory or literature. The answers will also provide interpretability to collaborators, funders, or even patients. All else equal, we seek to be able to explain how the model combines exposures to suggest that certain patients may be high risk and others may be low risk.

[4]But what if there is not a monotonic relationship due to functional form or just finite sample variation? Should we still estimate the low and high risk bins data-adaptively? I lean against it for simplicity but it remains an option.

As an exploratory aid to the interpretation of the mixture we provide functions to display the contribution (weight) of each exposure in the parametric OLS model. For all models we also show the average value of each exposure in each quantile of the mixture, and compare to the average value of each demographic variable to highlight potential confounding. Variables which are important contributors to the mixture will co-vary along with the mixture quantiles. Less important exposure variables will not be aligned with the mixture quantiles and therefore will tend to be stable around their mean across quantiles.

Complex exposure mixtures might require more effort to interpret, for example through model-agnostic interpretable machine learning such as accumulated local effects plots (Molnar 2019). Use of targeted predictive variable importance via Williamson et al. (2017)'s "vimp" approach may be a particularly useful ensemble interpretation method to integrate into the CV-TMLE procedure.

An attributable risk-style target parameter may be helpful to rigorously test for whether an individual exposure is incorporated into a mixture, and the directionality & relative potency of its impact.

$$\mathbb{E}[Y \mid f(A)] - \mathbb{E}[Y \mid f(a_i = 0, A_{-i})] \tag{3.2}$$

Or an alternative way to write this:

$$\mathbb{E}[Y_{f(A)}] - \mathbb{E}[Y_{f(a)} \mid a_i = 0] \tag{3.3}$$

Given a continuous exposure, the counterfactual intervention would set the given exposure to its lowest quantile (e.g. quartile or decile), while leaving the other exposures at their natural levels. If the exposure is not part of the mixture, then the risk difference will approximately zero and the p-value will be non-significant, possibly after multiple testing correction. If the exposure is part of the mixture then this population intervention measure (PIM) will provide an estimate of the direction and relative potency of its influence (Hubbard and M. J. van der Laan 2008). Among toxic exposures we would typically expect a negative direction from the population intervention measure. However, in some cases we may see a positive direction: for example, immunosuppressant exposures can be protective for childhood leukemia.

Another downside is that the method is sensitive to the contribution of the exposure within a specific quantile, which may not be representative of its overall effect. Consider if a given exposure influences the mixture at all quantiles with the exception of its lowest ($a = 0$) (e.g. due to a minimum concentration necessary for biological effect). In that scenario our focus on the initial quantile will lose substantial information. Perhaps instead we could use the omnibus nonparametric test of the causal null for continuous exposures as described in Westling (2020).

Can we estimate something along the lines of attributable risk to test for interaction as well? It will be important to design the analysis of interactivity to scale to a large number of exposures in high-dimensional data. Although we could have mixtures combining results

from only two exposures, such as the tobacco-related compounds nicotine and cotinine, we could reasonably include dozens or more exposure variables as part of a chemical mixture. And if we apply this method to genomic data we could have hundreds or even thousands of SNPs or methylation sites.

Similar to the attributable risk parameter, could we generate a dose-response curve for each individual exposure, either directly on the outcome or to characterize its influence within the mixture (E. H. Kennedy, Ma, et al. 2017; M. J. van der Laan, Bibaut, et al. 2018)? This could give us a granular visualization of an exposure's influence as a helpful complement the population intervention parameter. In this analysis, we might consider the predicted continuous mixture to be a sort of outcome variable of interest. We could then therefore conduct a typical varimpact analysis, where we loop over each exposure that is part of the mixture, discretize the individual exposure into quantiles, estimate the low and high risk levels, then evaluate that target parameter on the test set CV-TMLE style (pooling the test sets) where we condition on the other exposures and the adjustment variables, and our outcome variable is the continuous mixture. That would allow ranking the exposures in terms of their contribution to the mixture, if any, and assessing directionality & relative magnitude (loosely, due to the discretization of each exposure).

As an alternative option, the estimation of the mixture function $f(A)$ could use specific algorithms for interpretability reasons. For example, we might use parametric algorithms such as OLS, lasso, or MARS (Friedman et al. 1991) as a working model for the mixture function so that we can review the coefficients to interpret the contribution of individual variables. We would still benefit from the overall procedure that includes nonparametric confounder adjustment, a targeted correction term, and fair effect estimation through CV-TMLE.

Lastly, is there a potential to decompose the mixture estimation itself into parametric and nonparametric components? That could entail estimating a parametric component of the mixture that is highly interpretable (and also more amenable to targeting), followed by a nonparametric component on the residualized data, followed by the remainder of the procedure described above.

$$\mathbb{E}[Y \mid A, W] = \big(f(A) + g(A)\big) - \mathbb{E}[f(A) + g(A) \mid W] + h(W) \tag{3.4}$$

Where $f(A)$ is estimated parametrically (OLS, lasso, etc.) and g(A) summarizes the nonparametric remainder of the mixture function through backfitting.

Subgroups of exposures

Our method can be applied to distinct subgroups of exposures, which would allow the subgroups to be ranked by their ability to influence the outcome of interest. Groups of exposures that are causally related to the outcome should show the greatest risk difference (or ratio) between the high and low quantiles of the mixture exposure. In the most straight-forward approach these exposure subgroups would be derived from scientific knowledge. For example,

when analyzing toxic chemicals we might group exposures by their chemical class: PCBs, PBDEs, PAHs, etc.

If such exposure subgroups were not known a priori we could apply unsupervised clustering methods to the exposures to generate the subgroups observationally. Potential clustering algorithms might included k-means (Forgy 1965), hdbscan (Campello et al. 2013), or HOPACH (M. J. van der Laan and Pollard 2003). As long as those clustering methods did not incorporate the outcome variable they could be conducted outside of cross-validation. If supervised clustering were used (i.e. based on examining associations with the outcome variable), that clustering would need to be run on the training fold within cross-validation in order to yield unbiased risk estimates for the mixture quantiles. That would then also require aligning the clustering results across cross-validation folds so that they are comparable, because the unique clustering run within each of the training folds would possibly yield different cluster compositions, orderings, and even cluster counts.[5]

If exposure groups were the size of a single variable, this method would be very similar to the *varimpact*-style single variable analysis.

3.3 Data

We applied our method to simulated datasets with known relationships and to a childhood leukemia case-control study of chemical carcinogens measured through household dust collection. We used simulation designs and data from the National Institute of Environmental Health Sciences's (NIEHS) 2015 environmental mixtures workshop (Taylor et al. 2016).

NIEHS 2015 mixture workshop simulations

In 2015 the National Institute of Environmental Health Sciences held a scientific workshop titled "Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology Studies", [6] following up on their 2011 chemical mixtures workshop titled "Advancing Research on Mixtures: New Perspectives and Approaches for Predicting Adverse Human Health Effects". The 2015 workshop focused on analyzing two simulated datasets with a variety of proposed mixture algorithms, allowing the algorithms to be compared on the extent to which they created mixtures using the correct exposures and in the proper direction of risk (ibid.). We evaluated our methodology using these two simulations in order to provide benchmarking to alternative methods.

Simulation 1 was designed as a prospective cohort study with 500 observations, a continuous outcome, 7 continuous exposures, and 1 binary confounder. Exposures were log-normally

[5]We would especially prefer to avoid different numbers of clusters on each training fold estimation, as it would complicate the alignment of clusters across cross-validation iterations. Therefore we may want to prespecify the number of clusters, e.g. based on an unsupervised method run on the full data.

[6]`https://www.niehs.nih.gov/news/events/pastmtg/2015/statistical/index.cfm` and `https://factor.niehs.nih.gov/2015/8/science-mixtures/index.htm`

distributed.[7] There was no missing data, mediation, colliders, or unobserved confounding. X_1, X_2, X_7 contributed positively to the outcome, X_4, X_5 negatively, and X_3, X_6 did not contribute. X_1 was twice as potent as X_2 and X_5 was 4.5 times as potent as X_4. X_1, X_2 were competitively antagonistic with X_4, X_5, and were synergistic with X_7. X_4, X_5 were antagonistic with X_7. More details are provided in the simulation answers PDF.[8]

Simulation 2 was designed as a cross-sectional study with 500 observations, a continuous outcome, 14 continuous exposures, and 3 confounders (Z_1, Z_2 continuous and Z_3 binary). When Z_3 equalled 0 $X_4, X_6, X_{11}, X_{12}, X_{14}$ were positively related to the outcome. When Z_3 equalled 1 X_1, X_4, X_{11}, X_{14} were positively related to the outcome. There were no interactions between exposures. More detailed are provided in the simulation answers PDF. [9]

Application: chemical exposures in household dust

Following the evaluation by simulation, we then applied our method to a case-control study of childhood leukemia in which chemical exposures were measured by sampling household dust (Metayer, Colt, et al. 2013; Whitehead, Brown, et al. 2013; Bailey et al. 2015; Whitehead, Adhatamsoontra, et al. 2017). The dataset consisted of 583 observations of which 306 were cases and 277 controls. The 62 exposures of interest were grouped into 7 chemical classes based on scientific background knowledge: PCBs (4 exposures), metals (7), insecticides (21), tobacco (2), herbicides (7), PAHs (7), and PBDEs (14). There were 10 observed adjustment variables: family income, sex, age at study admission, Hispanic background, racial background, missingness indicator for racial background, number of previous residences, father's education, missingness indicator for father's education, and mother's education.

3.4 Results

We now demonstrate the results of the method using a combination of visualization and statistical reporting.

NIEHS 2015 mixture workshop simulations

NIEHS dataset 1

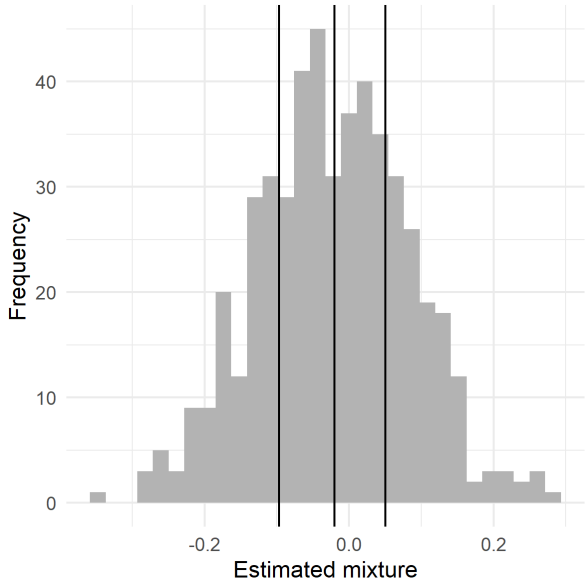
Results for the initial NIEHS 2015 simulation dataset are show below.

[7] Automate log transformation of exposures based on a test of normality on the raw data vs. log-transformed data, e.g. parametric KS-test?

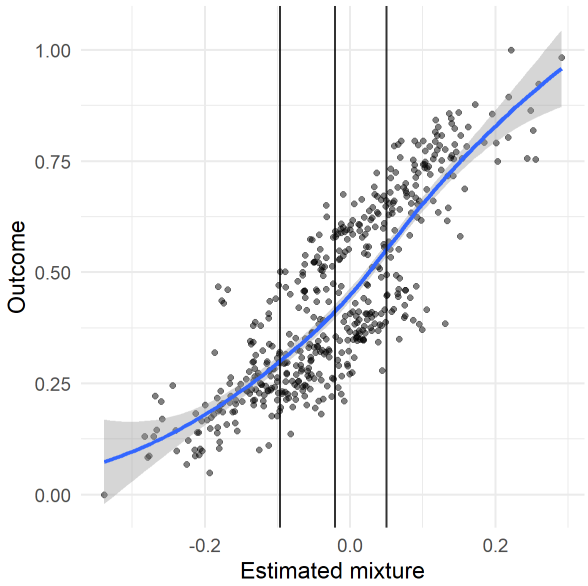
[8] `https://www.niehs.nih.gov/news/events/pastmtg/2015/statistical/simulated_dataset_1_answers.pdf`

[9] `https://www.niehs.nih.gov/news/events/pastmtg/2015/statistical/simulated_dataset_2_answers.pdf`

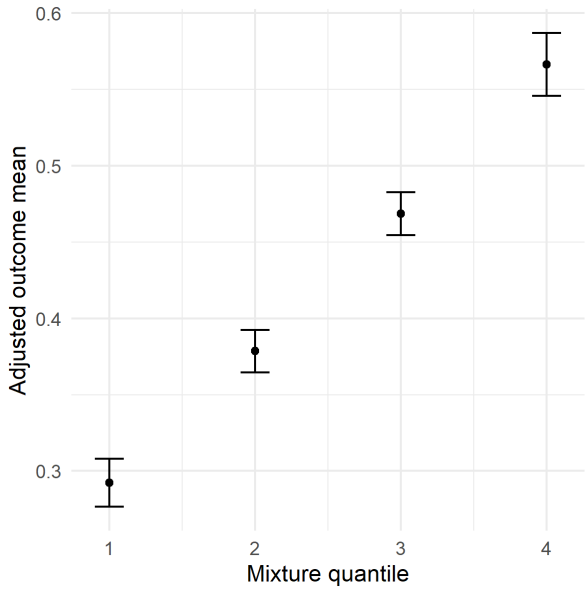
(a) Distribution of the estimated mixture, with quartile thresholds shown as vertical black lines.



(b) Unadjusted risk estimate from lowess smooth, and with quartile thresholds shown as vertical black lines.



(c) Adjusted risk and 95% confidence intervals estimated for each mixture quartile using cross-validated targeted learning.



(d) Distribution of standardized adjustment variables (potential confounders) across quantiles of the mixture.

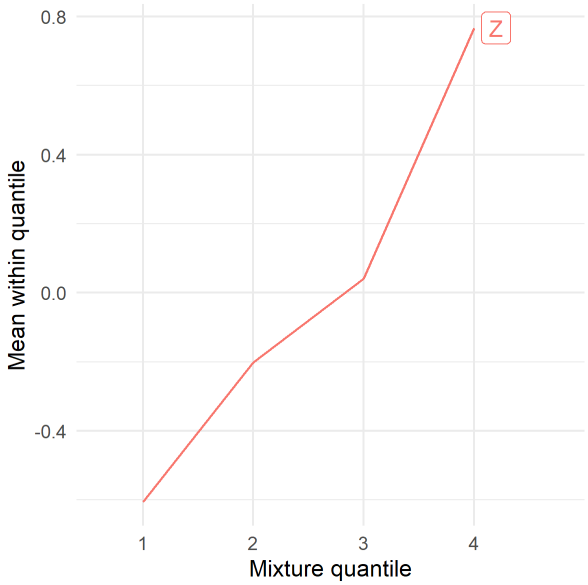


Figure 3.1. Mixture analysis on NIEHS simulated dataset 1

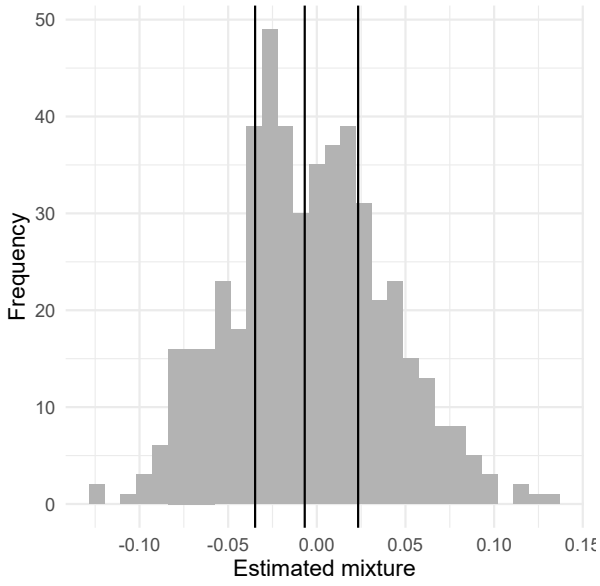
Table 3.1. Summary of mixture analysis for dataset 1

	Mixture quantiles			
	Q1. [-0.348, -0.082)	**Q2.** [-0.082, 0.003)	**Q3.** [0.003, 0.081)	**Q4.** [0.081, 0.345]
Mixture	-0.149	-0.039	0.041	0.147
Outcome	0.230	0.355	0.474	0.685
Exposures				
X1	-0.985	-0.449	-0.054	0.542
X2	-0.476	-0.143	-0.044	0.230
X3	-0.954	-0.384	-0.071	0.480
X4	-0.147	-0.119	-0.002	0.168
X5	0.288	0.100	-0.241	-0.685
X6	0.049	-0.010	-0.161	-0.370
X7	-1.163	-0.372	0.019	0.474
Adjustment				
Z	-0.622	-0.218	0.056	0.783

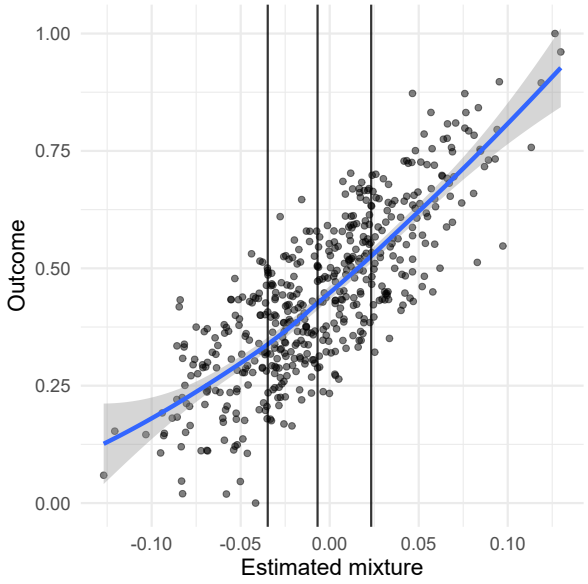
NIEHS dataset 2

Results for the second NIEHS simulated dataset are shown below.

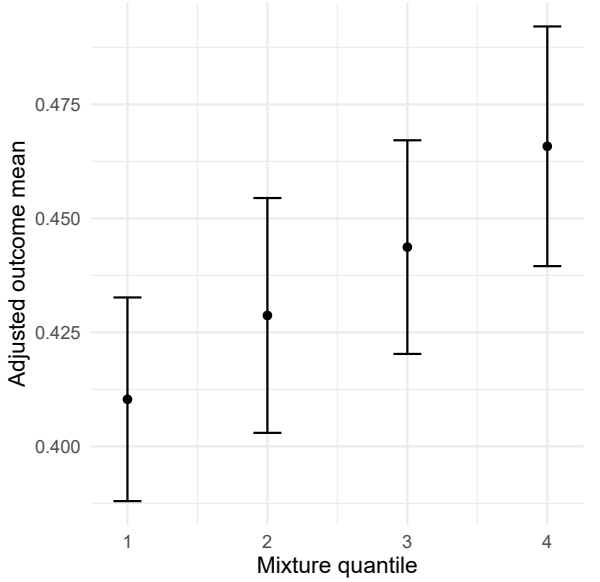
(a) Distribution of the estimated mixture, with quartile thresholds shown as vertical black lines.



(b) Unadjusted risk estimate from lowess smooth, and with quartile thresholds shown as vertical black lines.



(c) Adjusted risk and 95% confidence intervals estimated for each mixture quartile using cross-validated targeted learning.



(d) Distribution of standardized adjustment variables (potential confounders) across quantiles of the mixture.

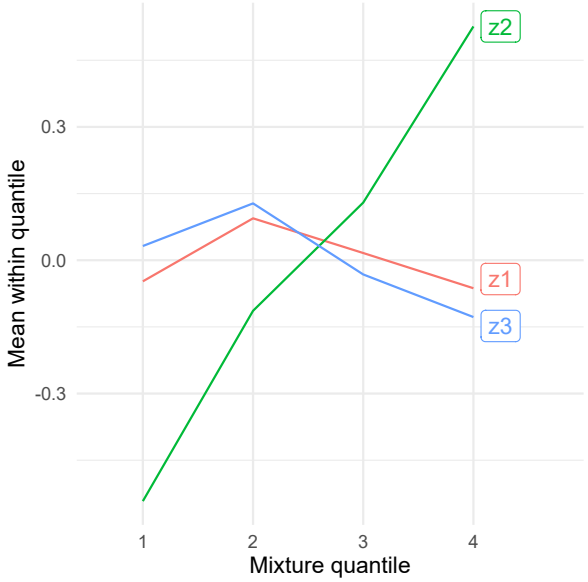


Figure 3.2. Mixture analysis on NIEHS simulated dataset 2

Table 3.2. Summary of mixture analysis for dataset 2

	Mixture quantiles			
	Q1. [-0.12696,-0.03481)	**Q2.** [-0.03481,-0.00682)	**Q3.** [-0.00682, 0.02330)	**Q4.** [0.02330, 0.12979]
Mixture	-0.060	-0.022	0.009	0.051
Outcome	0.278	0.376	0.484	0.622
Exposures				
x1	0.719	1.026	1.090	1.258
x2	-2.298	-2.211	-2.115	-1.859
x3	0.444	1.089	1.556	2.209
x4	1.457	2.099	2.591	3.243
x5	-0.468	0.266	0.798	1.551
x6	0.089	0.666	1.095	1.713
x7	1.019	1.190	1.398	1.678
x8	1.936	2.584	2.851	3.423
x9	0.984	1.212	1.424	1.674
x10	2.644	3.028	3.251	3.643
x11	4.645	5.036	5.308	5.765
x12	0.337	0.397	0.552	0.637
x13	0.454	0.473	0.613	0.682
x14	0.724	1.227	1.439	1.857
Adjustment variables				
z1	-0.048	0.094	0.016	-0.063
z2	-0.542	-0.114	0.130	0.526
z3	0.032	0.128	-0.032	-0.128

Chemical exposures in household dust

As an application to a scientific mixture problem, we analyzed the household dust dataset using exposure groupings based on chemical classes. Those 7 chemical classes are: herbicides, insecticides, PAHs, PCBs, PBDEs, metals, and tobacco. The outcome variable is case-control status, and we have 227 observations with complete data (112 cases, 115 controls). There were 10 adjustment variables as described earlier (family income, sex, race, maternal age, etc.).

We created each mixture using only the exposures in that class, adjusted for covariates, applied the mixture and quantiles to the test data, and estimated the counterfactual effect of all observations being in the high quantile vs. the low quantile. Exposure groups were then ranked based on the statistical significance of that average treatment effect. We reported the Benjamini-Hochberg multiple testing adjusted p-value in addition to the unadjusted

p-value (Benjamini et al. 1995). During mixture estimation we did not adjust for other exposures beyond the group being analyzed, due to the small sample size and large number of exposures.

Table 3.3. Ranking of exposure groups based on mixture impact

	Group	ATE	P-value	Std. err.	Adj. p-value	95% CI
1	tobacco	-0.426	0.000	0.040	0.000	(-0.504, -0.349)
2	pbdes_2	-0.380	0.000	0.046	0.000	(-0.469, -0.29)
3	pah	-0.330	0.000	0.052	0.000	(-0.432, -0.228)
4	insecticides	-0.162	0.001	0.049	0.001	(-0.258, -0.066)

3.5 Discussion

Experimental design

This work bears similarity to experimental design in randomized trials with multiple dimensions of treatment, or continuous factors. The design of experiments (DOE) for factorial trials and corresponding response surface methodology (RSM) show remarkable correspondence to the problems we face in exposure mixtures. In fact, factorial designs are the preferred theoretical experiment for mixture analysis: randomization ensures no confounding in expectation, and interaction effects across dimensions of treatment can be analyzed by inclusion of interaction terms in linear regression analyses. Moreover, the study of the design of experiments includes a subfield termed "mixture experiments" that are particularly relevant to our work here in observational mixtures (Goos et al. 2011; Lawson 2014; Myers et al. 2016). We leave to future research the exploration of integrating observational and experimental mixture designs, particularly with regard to blending effects (interactions) and response trace plots (dose-response curves).

There remains a key benefit to using our mixture methodology even given a lack of confounding: the nonparametric ensemble estimation of the mixture function will automatically capture interactions and nonlinearities of the treatment dimensions. Then the mixture can be interpreted using the analysis methods we have described, to understand the contributions of each dimension to the overall experimental response surface. To our knowledge nonparametric ensemble machine learning has not yet been applied to response surface analysis of experimental designs.

Factorial experimental designs are also a way to validate the mixture estimation function. In the case of childhood leukemia that validation could be conducted in a mouse model, and toxicologists often analyze mixtures experimentally using bacterial and other models. However, a factorial design is not quite the ideal due to the potential for sparsity in the mixture. One or more exposures may not contribute to the mixture, and if that is the case, all replication combinations of treatments in which those exposures vary do not contribute

additional information about the mixture function. And even if all exposures do contribute to the mixture, there will be variance in the potency across the exposures. Thus if we acknowledge the likely differential importance of exposures towards the mixture function, we will benefit from a "random search" style design where sample points randomly from the response surface (Bergstra et al. 2012). That will ensure that even if certain dimensions do not vary on the response surface we will still test variation in the values measured along the other dimensions.

Missing data

As in any causal inference exercise, we would like to take into account that missingness in the exposure variable(s) may be biasing our estimation. In a single exposure analysis we would create a missingness indicator for the exposure, then model that exposure using the adjustment variables to estimate the probability of missingness for all observations. In the same vein as propensity score weighting, we would then conduct our analysis (graphical computation, logistic regression, TMLE, etc.) but weighting the non-missing observations by the inverse probability of being non-missing. That would up-weight non-missing observations to the extent that they are representative of the observations that are missing the exposure. This procedure is called inverse probability of censoring weighting (IPCW).

Missing data in the exposure variables poses more of a problem for mixtures than single-exposure analyses. On the one hand, as the exposure vector becomes increasingly high dimensional we would expect that the probability that all observations have a missing value in at least one exposure would tend towards 1. In other words, a complete case analysis will be increasingly detrimental to statistical power as we analyze more exposures. And on the other hand, we can't create a separate missingness models for each exposure - we would have to create a combined missingness model that re-weights the complete cases to best describe the incomplete cases. So we increasingly lose the information in any exposures that have missingness. The question then becomes: is there a way to use the exposure measurements in the incomplete cases in order to improve power over an IPCW complete case analysis? For example, is it beneficial to impute the exposures during mixture estimation, even though we do not want to rely on the imputation model being correctly specified? We leave these questions to future research.

With regard to missingness in the adjustment variables, that is less of a concern because standard practice will suffice. We will include missingness indicators for each adjustment variable that has any missing values, then impute the missing cells as best we can. This procedure is recognized to be adequate in general (Blake et al. 2020).

Future work

The next step in the evaluation of this proposed methodology is to conduct a thorough benchmarking approach to existing key algorithms, as in Lazarevic, Knibbs, et al. (2019).

Another item for future exploration is the pursuit of "mixtures of mixtures". That is, once we identify mixtures of groups of exposures, can we combine those mixtures into super-mixtures that incorporate interaction effects across the constituent mixtures? That might be accomplished through nested CV-TMLE, or simply by analyzing all exposures at once.

As a complementary item, we also hope to explore automatic clustering of exposure variables into subgroups, for situations in which scientific knowledge of subgroups is not available. Additional tasks include supporting case-control weights, and providing statistical power analyses for the design of new studies.

Given the generality of the method, we will consider its utility for the analysis of factorial or fractional factorial experimental designs. In that scenario confounding is not an issue but the projection of the possibly high-dimensionality treatment vector onto a unidimensional mixture, and the ensuing interpration analysis, may be beneficial.

3.6 Conclusion

It is important to recognize the generality of the proposed method for mixture problems. While we focused on environmental health, our method could be applied to nutritional exposures (e.g. via food frequency questionnaires), pharmaceutical prescriptions in an electronic health record, maternal comorbidities on pregnancy outcomes, neighborhood characteristics on social outcomes, molecular structures on protein characteristics, SNPs on diseases, etc. Phrases in a paragraph (Fong et al. 2016), pixels in an image, or financial stocks in a market might also be converted into multiple exposures that can be analyzed as mixtures. We see mixture-based analyses as relevant to any problem in which a vector of treatments or exposures is measured and whose effect on an outcome is desired.

What is defined as a single-exposure versus a mixture may also be in the eye of the beholder. If the "single" exposure is instead viewed as an "omnibus treatment", it may be possible to decompose that treatment into multiple components that can be analyzed as a mixture, provided that the components are not perfectly correlated.

3.7 Appendix

Formal argument for double robustness

We present a formal argument for establishing that our method for estimation of $\mathbb{E}[Y \mid A, W] - \mathbb{E}[Y \mid W]$ is double robust.

Suppose we asymptotically solve for a given $\theta(W)$ and $g(A \mid W)$ a class of estimating functions $0 = \mathbb{E}_0 S_{g,j}(A \mid W)(Y - \theta(W) - m(A, W))$ indexed by a choice of $S_{g,j}(A \mid W)$ satisfying $\mathbb{E}[S_{g,j}(A, W) \mid W] = 0$. For example we have a whole set of basis functions $\phi_j(A, W)$ and we center them as $\phi_j(A, W) - \mathbb{E}_g[\phi_j(A, W) \mid W]$. We solve this under constraint that $m(A, W)$ satisfies $\mathbb{E}_g[m(A, W) \mid W] = 0$ (by only including basis functions $\phi_j - \mathbb{E}_g[\phi_j \mid W]$).

One could succeed in solving such equations (asymptotically) by:

1. First estimate $\theta(W)$, estimator θ_n of $\mathbb{E}[Y \mid W]$.
2. Estimate g_n of $g(A \mid W)$; we do this implicitly by centering basis functions.
3. Use $\theta_n(W)$ as offset and now run regression of Y on $\phi_j - \mathbb{E}_{g_n}[\phi_j \mid W]$, $j = 1, \ldots, J$. This regression fit is then $\theta_n(W) + m_n(A, W)$. For example, one could use lasso with these basis functions ϕ_j (e.g. HAL by letting ϕ_j be the HAL basis). By undersmoothing HAL one would solve:

$$P_n(\phi_j - \mathbb{E}_{g_n}[\phi_j \mid W])(Y - \theta_n(W) - m_n(A, W)) = o_P(n^{-1/2}) \tag{3.5}$$

We now want to show why this approach has a robustness property in the sense that m_n will be consistent if either g_n or θ_n is consistent for its desired target $g_0(A \mid W)$ and $\theta_0 = \mathbb{E}[Y \mid W]$.

Let g and θ represent the limits of our estimators of θ_n and g_n and let m be the limit of m_n. Then m solves:

$$0 = P_0 S_{g,j}(A, W)(Y - \theta(W) - m(A, W)) \tag{3.6}$$

for all j, i.e. a huge collection of basis functions. In addition, m is forced to satisfy $\mathbb{E}_g[m(A, W) \mid W] = 0$ w.r.t. the g we used (since we only use basis functions $\phi_j - \mathbb{E}_g[\phi_j \mid W]$).

Proof of robustness

Take conditional mean, given A, W:

$$0 = P_0 S_{g,j}(A, W)(\mathbb{E}[Y \mid A, W] - \theta(W) - m(A, W))(\cdot) \tag{3.7}$$

Suppose $\theta(W) = \theta_0(W)$. Then this becomes:

$$0 = P_0 S_{g,j}(A, W)(\mathbb{E}[Y \mid A, W] - \mathbb{E}[Y \mid W] - m(A, W)) \tag{3.8}$$

for all j. This implies that:

$$m(A, W) = \mathbb{E}[Y \mid A, W] - \mathbb{E}[Y \mid W] + h(W) \tag{3.9}$$

for some h. However, we also know that $\mathbb{E}_g[m(A, W) \mid W] = 0$. That implies $h(W) = 0$, so we have:

$$m(A, W) = \mathbb{E}[Y \mid A, W] - \mathbb{E}[Y \mid W] \tag{3.10}$$

Suppose now that $g = g_0$. Consider $(\cdot)$ again.
Firstly note that it equals:

$$0 = P_0 S_{g_0,j}(A, W)(\mathbb{E}[Y \mid A, W] - m(A, W)) \tag{3.11}$$

since $\theta(W)$ cancels due to $\mathbb{E}(S_{g_0,j}(A, W) \mid W] = 0$.
Now, we can also add the true θ_0 back in:

$$0 = P_0 S_{g_0,j}(A, W)(\mathbb{E}[Y \mid A, W] - \theta_0(W) - m(A, W)) \tag{3.12}$$

using the same trick. But that means that:

$$m(A, W) = \mathbb{E}[Y \mid A, W] - \theta_0(W) + h(W) \tag{3.13}$$

for some h(W). However, we also know that $\mathbb{E}_{g_0}[m(A, W) \mid W] = 0$, which thus yields $h(W) = 0$ and therefore:

$$m(A, W) = \mathbb{E}[Y \mid A, W] - \theta_0(W) \tag{3.14}$$

This completes the proof.

Parametric mixture with parametric confounder adjustment

In the simplest mixture estimation method we would use ordinary least squares (OLS) to estimate the linear combination of exposures that best predict the outcome, including linear adjustment for confounding via W.

$$E[Y \mid A, W] = \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_k A_k + \beta_* W \tag{3.15}$$

OLS-based estimation provides some causal defensibility because we are adjusting for confounding in W, compared to no adjustment at all. It also allows for interpretability of the individual exposure contributions through examining the coefficient signs and magnitudes (dependent upon scale though). Limitations include the strong potential for residual confounding due to the restrictive linear adjustment, lack of detection of interaction between exposures, and lack of finite sample performance because the bias-variance tradeoff is optimized for predicting the outcome rather than optimizing the mixture estimation.

Parametric mixture with nonparametric confounder adjustment

A key improvement is to expand the mixture estimation to support semi-parametric regression. Such an estimator would allow for nonparametric adjustment for confounding (missing from the ordinary least squares version) while keeping the exposure mixture separate from the confounder adjustment. One version might be in the form of a parametric beta vector with a learned nonparametric f function:

$$E[Y \mid A, W] = \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_k A_k + f(W) \tag{3.16}$$

An important limitation of this approach though, is that we are not detecting or allowing interactions between exposures to modify their combined effect. Yet that is a key goal of mixture analysis: understanding how combinations of exposures may influence health outcomes due to synergistic or antagonistic interactions. Thus we may want to expand the parametric form to support two-way interactions for example:

$$E[Y \mid A, W] = \beta_1 A_1 + \cdots + \beta_k A_k + \beta_{k+1} A_1 A_2 + \beta_{k+2} A_1 A_3 + \ldots \beta_{k*} A_{k-1} A_k + f(W) \tag{3.17}$$

As the number of exposures grows, those two-way interactions will grow exponentially. For example, with 15 exposures there would be 15 * 14 / 2 = 105 two-way first-order interaction terms. That growth in complexity would presumably necessitate that some sort of sparsity, penalization, or feature selection is applied. Dividing exposures into subgroups and restricting two-way interactions to exposures within a subgroup could be another way to maintain power.

Scaling of exposures

When using a linear method for estimating the exposure mixture function, such as least squares, partial least squares, or weighted quantile sum regression, and reviewing the estimated parameters (weights) to help interpret the composition of the estimated mixture, the

scale of the exposures determines the scale of the coefficients. In order for the coefficient magnitudes to be an interpretable, scale-free estimate of which exposures most contribute to the mixture we need the exposures to be on comparable scales. One method for this has been to discretize the exposures into quantiles, e.g. quartiles or deciles, possibly based on histogram scaling (Rozenholc et al. 2010). Another option is standardizing the exposure means and standard deviations to be 0 and 1 respectively, possibly after a log-transformation.

Chapter 4

Clinical prediction modeling using electronic health records

> The omission of prediction from the major goals of basic medical science has impoverished the intellectual content of clinical work, since a modern clinicians main challenge in the care of patients is to make predictions.
>
> Alvan Feinstein, 1983

4.1 Introduction

Chest pain is the second leading reason for emergency department visits (Rui et al. 2016) and is commonly identified as a leading driver of low-value health care. Workup protocols in patients with chest pain are designed to diagnose the potential for major adverse cardiac events (MACE). Missed diagnoses of MACE can be cause for medico-legal action, which may encourage conservative testing without health benefit. Accurate identification of patients at low risk of MACE is important to improve resource allocation and reduce overtreatment (Amsterdam et al. 2010). Risk scores aim to identify patients eligible for early discharge, avoiding additional stress testing and cardiac imaging that is unlikely to be of benefit (Greenslade et al. 2018). The primary biomarkers used for initial triage are elevated cardiac troponin, a sensitive marker of cardiac injury measured serially, and repeated electrocardiograms.

Previous work has focused on the development and validation of additive risk scores as decision aids for risk stratification. Such risk scores examine a small number of biomarkers and demographics, summarize those predictors into qualitative levels, and use a weighted

sum to allocate patients into risk categories. Standard risk scores are HEART (History, ECG, Age, Risk factors and Troponin) and EDACS (Emergency Department Assessment of Chest Pain Score - Than, Flaws, et al. 2014). HEART is most commonly used in North America, although EDACS has similar performance characteristics (Mark et al. 2018). Effective risk scores will stratify patients across risk levels such that the qualitative "low risk" group will have sufficiently low risk of short-term MACE that those patients can be discharged without additional workup. An ineffective or ill-calibrated risk score would underestimate the risk in the "low risk" group and lead to an overly optimistic early discharge policy that results in increased future MACE. But given multiple risk scores that are well-calibrated, scores with improved discrimination could theoretically result in a larger percentage of low-risk patients.

Background and Objectives

It remains debated whether machine learning methods can exhibit statistically and substantively significant benefits for risk prediction compared to logistic regression, decision trees, or additive risk scores (Goldstein, Navar, and Carter 2016; Goldstein, Navar, Pencina, et al. 2016). A recent meta-analysis, for example, did not find systematic benefit from machine learning in comparison to logistic regression (Christodouloua et al. 2019). Yet there is also optimism about the potential for artificial intelligence methods in medicine (J. He et al. 2019) in general, as well as cardiology specifically (Johnson et al. 2018).

Building on Mark et al. 2018, we sought to assess the performance of machine learning (ML) methods at predicting MACE among emergency department patients with chest pain. Could ML improve upon existing validated risk scores through a more complex integration of predictors that can better estimate MACE risk? To what extent is hyperparameter optimization necessary to achieve strong ML performance?

Our clinical objective was to maximize the pool of low-risk patients that are accurately predicted to have less than 0.5% MACE risk and may be eligible for reduced testing. The primary threshold of 0.5% risk has previously been identified as an acceptable risk by a majority of emergency physicians for early discharge (Than, Herbert, et al. 2013). Using a risk of 0.5% as the test threshold will inherently lead to a negative predictive value of greater than 99.5%, provided that the risk prediction is well-calibrated in the target population. We also examined secondary thresholds of 1.0% and 2.0%.

A reasonable assessment of ML performance could only be made in comparison to realistic alternative options. We compared ML performance to simpler indicators of risk: key biomarkers (troponin, electrocardiogram), validated clinical risk scores (History, ECG, Age, Risk factors and Troponin [HEART] and Emergency Department Assessment of Chest pain Score [EDACS]), decision trees, and logistic regression.

If machine learning can demonstrate improved discriminative performance compared to logistic regression and related methods, along with appropriate calibration, its next hurdle for adoption is to provide interpretability. Clinicians may be willing to forgo maximum predictive accuracy for the sake of understanding how individual predictors influence the output of the algorithm. With analytical effort it may be possible to provide sufficient

interpretability for clinicians to accept the complication of machine learning and the benefit of the (potentially) improved predictive accuracy. To facilitate interpretation we explained the models through prediction-based variable importance ranking and accumulated local effect visualization. If simpler algorithms remain preferred, the ML results can at least approximately the best achievable performance, and so serve as benchmark standards when considering more restrictive algorithms.

Certain analytical characteristics would be important to arrange in order for ML to potentially improve upon simpler options. First, it was important to extract a broader set of granular predictor variables than were used by existing scores. Extensive predictor sets give ML the potential to capture interactions and nonlinear relationships that are missed by linear or additive approaches, perhaps relevant only to certain subgroups of patients. Further, ML may statistically identify novel predictors that have been missed by existing scores or the broader literature, or whose predictive impact was too small, or in too complex a form to be detected by non-ML methods (or may have been inadequate for a given sample size). And the expansion of electronic health records (EHRs) makes broader covariate collection more feasible and relevant than was possible prior to EHRs, while also facilitating more granular measurement of variables (E. H. Kennedy, Wiitala, et al. 2013).

Next, it was important for those variables be measured on a fine-grained scale, which gives ML the opportunity to detect novel cut-points or thresholds that improve performance. Variables should be kept as their original continuous measurements rather than dichotomized or discretized into qualitative levels (Senn 2005). For example, a predictor such as BMI loses substantial information when it is dichotomized into an indicator of high-BMI or the absence of high-BMI. A single threshold chosen for for that dichotomization may not be optimal for certain subgroups or regions of risk. One of the benefits of ML is that it can identify thresholds in a data-adaptive way, allowing it to better approximate unknown or ill-understood physiological processes. That said, very high cardinality variables can result in overfitting and slow down the training of certain machine learning algorithms, such as decision trees, that test each unique value as a possible subgroup splitting threshold. It may be beneficial to reduce the cardinality of granular continuous variables through histogram binning that scales with the dataset size, e.g. sample size / 1000 or log(sample size) * 10.

4.2 Data and Methods

Source of data

Our study was sourced from the electronic health record (EHR) of 21 emergency departments within Kaiser Permanente Northern California, an integrated health care delivery system.

Participants

All adult patients were retrospectively included if they had received cardiac troponin testing in the emergency department and either presented with a chief complaint of chest pain or chest discomfort, or whose ED physician had assigned them a primary or secondary ICD-coded diagnosis of chest pain. The later inclusion criterion is important because patients may complain of "anginal equivalents" (such as shortness of breath) in lieu of overt chest pain (Amsterdam et al. 2010). That initial inclusion pool had a 60-day MACE rate of 8.0%. Patients were excluded if they had a MACE diagnosis in the ED or within 30 days prior to ED visit, alternative non-ACS diagnoses at index visit (e.g. pneumonia, pneumothorax, or traumatic injury), could not be tracked due to lack of active health plan membership during the study (except in cases of death), or with troponin I $>$ 99th percentile upper limit of normal. Patients were excluded if their smoking status was unknown, which was viewed as a key marker of low-quality data. The final study cohort consisted of 116,764 patients with a 60-day MACE incidence of 1.88%. The troponin assay used during the study period was the AccuTnI+3 (Beckman-Couleter, Brea, CA, USA).

Outcome

Our primary outcome was cumulative MACE incidence within 60 days of the index visit. We defined MACE as myocardial infarction, cardiogenic shock, cardiac arrest, or death.

Predictors

We used a total of 74 predictors sourced from the electronic health record, including vitals, labs, history, qualitative interpretation of ECG imaging, regular expression-based extraction of features from clinical notes, demographics, and missingness indicators (20). These predictors are detailed in Table 4.4.

Missing data

Missingness rates for each predictor are listed in Table 4.4. We created missingness indicators for each predictor, which marked the observations that were missing a value. That matrix of missingness indicators was analyzed for perfect collinearity, and duplicate indicators were dropped.

Missing predictor values were imputed by factorizing the raw data matrix with generalized low-rank models (GLRM) (Schuler et al. 2016; Udell et al. 2016). GLRM is a generalization of principal component analysis and matrix completion methods and is designed for mixed type data frames that include continuous, categorical, ordinal, and binary variables. GLRM decomposes (factorizes) the original data frame into an X matrix of reduced components and Y matrix of archetypes, including possible penalty terms that can induce sparsity (L1) or simply denoise (L2 or quadratic). Multiplying these two factor matrices reconstructs the

original data frame, imputing any data entries with missing values. The method used for missingness provides few constraints on the resulting fit and also permits prediction from future data with missing values.

The GRLM hyperparameter settings were chosen through a grid search in which each model was trained on 75% of the data and evaluated on the remaining 25% for accuracy at reconstructing the original observed data matrix. Missingness indicators were not included in the GRLM imputation analysis. Our final GLRM settings were: 50 components, quadratic regularization on X with weight 4, and L1 regularization on Y with weight 24. Cells with missing data were then replaced with the reconstructed data matrix from GLRM using the optimal settings.[1]

GLRM imputation greatly increased the number of unique values (cardinality) for continuous variables,which would have a negative performance impact on tree-based algorithms that test every unique value for a potential split. To avoid that performance drop, we using penalized histogram binning to bin imputed predictors with high cardinality into up to 200 unique values (Rozenholc et al. 2010).

Multiple imputation was not necessary because our scientific goal was to characterize predictive performance for the unimputed outcome variable, rather than to estimate statistical parameters for covariates that were imputed, such as linear regression coefficients (Wang et al. 1992; Steyerberg 2009).

Prediction algorithms

Dozens if not hundreds of other prediction algorithms would be possible to evaluate, but computational time limitations forced us to choose a finite set with reasonable performance expectations. We chose well-known prediction algorithms that have shown strong performance in prior research, including both linear and decision tree-based estimation. The tree-based prediction algorithms were random forest (Breiman 2001), extreme gradient boosting (XGBoost) (Chen and Guestrin 2016), and Bayesian additive regression trees (Chipman et al. 2010). The linear prediction algorithms were generalized additive models (T. J. Hastie et al. 1990) using thin plate splines (Wood 2003), and lasso (Tibshirani 1996).

Splines have shown competitive performance with tree-based algorithms in prior clinical prediction work due to their ability to identify non-linear, but smooth patterns (Austin 2007). The lasso algorithm (or its generalization the elastic net) is a helpful test of sparsity in the covariates, and a faster & more nuanced variable selection method than best subset or stepwise selection (T. Hastie et al. 2017). Better performance for lasso compared to logistic regression would indicate that feature selection could be helpful for other algorithms, while equal performance could indicate that the extraction of predictors from the EHR was overly restrictive and should be broadened.

[1]Here X refers to the reduced components after GLRM transformation, and Y refers to the complementary matrix that transforms those components back to the original covariate space. It does not refer to the outcome variable.

Benchmarks

When evaluating complex algorithms it is important to contextualize their performance by comparing to simpler alternative approaches or benchmarks. If the benchmark algorithms can achieve similar performance then the extra complexity of the statistical machine learning algorithms may not be worthwhile. The improvement of a novel prediction method over standard benchmarks is known as the *skill* of the prediction method (Brier 1950; F. Sanders 1963; Murphy et al. 1977). In clinical prediction the primary alternatives to statistical machine learning are relatively inflexible fits, which include logistic regression, ordinary least squares, individual decision trees, and stratification on key clinical covariates. We tested each of these options, where key covariates were defined as peak troponin, qualitative ECG reading, EDACS score, and HEART score. As a complement to stratification on different subsets of key covariates, we also evaluated logistic regression and decision trees when restricted to these key covariates.

Stacked ensembling

When comparing a variety of algorithms an initial choice is to use cross-validation to select the algorithm with the best out-of-sample performance. A more nuanced decision would be to consider a weighted average of multiple algorithms - creating a team of algorithms whose contribution to the prediction is based on optimizing out-of-sample performance on a certain statistic. That is the nature of stacked ensembles (Wolpert 1992; Breiman 1996), sometimes referred to as the Super Learner algorithm (M. J. van der Laan, Polley, et al. 2007). Rather than restrict our prediction machine to a single algorithm, we create a weighted average across all tested algorithms, and select weights based on an optimization goal so that they minimize a chosen performance statistic on test data. We chose to optimize the Brier score (i.e. mean-squared error) in our ensemble, using convex weights based on a non-negative least squares meta-learner. Optimizing on Brier score includes a focus on both discrimination and calibration for the ensemble (Murphy et al. 1977). A convex combination of algorithm weights ensures that predictions fall within the convex hull of the constituent learners, while also inducing sparsity - i.e. algorithms can have zero weight.

Hyperparameter tuning

Prediction algorithms often have a multiple hyperparameter settings that adjust the estimation procedure in different ways. Those hyperparameters are not estimated from the data, but rather must be specified a priori by the analyst. While software implementations will typically provide a default value for each hyperparameter, there is no reason to believe that the default values are effective for the current dataset. Customizing the hyperparameter configuration to the current dataset can allow the algorithms to adapt to the available sample size, number of predictor variables, measurement error in the predictors, sparsity in predictor relevance, and correlation structure of the predictors. Hyperparameters are often

chosen by fitting the algorithm with different configurations and selecting the configuration that maximizes accuracy on held-out data, such as through cross-validation. The benefit of hyperparameter tuning is believed to vary by algorithm, which is referred to as the *tunability* of the algorithm (Probst, Boulesteix, and Bischl 2019). Random forest, for example, is believed to work well with default hyperparameters but also can benefit from hyperparameter tuning, particularly to reduce overfitting (Segal et al. 2011; Probst, M. N. Wright, et al. 2019).

Hyperparameter tuning is inherently a computationally intensive process, as it involves fitting the algorithms many different times, and varies based on the number of hyperparameters (dimensionality) as well as number of the unique values tested for each hyperparameter (resolution). Further complexity is involved if one considers that some hyperparameters may be more important than others for a given algorithm. Given the role of hyperparameters in modifying the performance of prediction algorithms, caution is warranted when generalizing algorithm performance characteristics from individual studies (e.g. algorithm X outperforms algorithm Y), particularly when hyperparameters are left at their default values and therefore are not customized to the given dataset.

For this work we adopted a hyperparameter tuning approach using *nested ensembling*. Much as using a weighted ensemble of different algorithms may be preferable to selecting the single best-performing algorithm, using a weighted ensemble of hyperparameter settings for a given algorithm may yield improved performance compared to selecting a single set of hyperparameters. With that concept in mind we created small grids of hyperparameter configurations and estimated a SuperLearner ensemble for a given algorithm in which the ensemble weights selected the hyperparameter settings that maximized out-of-sample performance. This ensemble of hyperparameter settings could potentially rely on a single configuration due to the sparsity induced by the convex combination, or the optimization could distribute the weighting across multiple configurations if such a weighting improved performance over a single selected configuration. Another benefit of the nested ensembling is that it limits the number of learners that are analyzed in the outer SuperLearner ensemble, which can conserve power and mitigate overfitting in the meta-learning process (i.e. allocation of weights in the convex combination).

We used the ensemble hyperparameter tuning approach for random forests, xgboost, and individual decision trees. The random forest grid consisted of 9 configurations: minimum node size $\in \{5, 20, 60\}$ $\times$ covariates sampled $\in \{4, 8, 16\}$ [2]. The xgboost grid consisted of 8 configurations: number of trees $\in \{250, 1000\}$ $\times$ maximum tree depth $\in \{2, 4\}$ $\times$ shrinkage $\in \{0.05, 0.2\}$. The decision tree grid consisted of 12 configurations: complexity parameter $\in \{0, 0.01\}$ $\times$ minimum split $\in \{10, 20, 80\}$ $\times$ maximum tree depth $\in \{10, 30\}$.

[2]The number of covariates sampled (i.e. mtry) was based on the formula: floor($\{0.5, 1, 2\} \cdot \sqrt{p}$) where p is the total number of covariates.

Evaluation

We evaluated alternative options for risk prediction based on their discrimination, calibration, and clinical utility. Nested cross-validation with 5 folds was used to conduct the discrimination and calibration analyses. While bootstrap estimation has been promoted for evaluation of clinical prediction models, recent work has shown the bootstrap can be biased for evaluating the performance of highly adaptive ML algorithms estimators such as random forests (Benkeser, Petersen, et al. 2019).

Discrimination

We chose area under the precision-recall curve (PR-AUC, also known as average precision) as our primary performance metric for evaluating discrimination, because it highlights performance differences that may be missed by ROC-AUC with imbalanced data (Cook 2007; Saito et al. 2015). We included area under the receiver operating characteristic curve (ROC-AUC or the concordance statistic) as our secondary performance metric, which remains highly popular and interpretable (Janssens et al. 2020). As an exploratory metric we also estimated the adjusted Brier score (index of prediction accuracy) which integrates discrimination and calibration into a single metric (Kattan et al. 2018). We did not conduct a reclassification analysis due to recognized limitations (Hilden et al. 2014; Kerr et al. 2014; Leening et al. 2014; Pepe et al. 2015).

Calibration

Our clinical use case was centered on a risk threshold of 0.5% to classify patients as "low risk" in order to qualify for early discharge. Because of that scientific goal, it was especially important to compare the model's predicted risks to the observed risks, i.e. its *calibration* (Lichtenstein et al. 1981) - also known as reliability (Brier 1950; Murphy et al. 1977), external validity, trustworthiness, or external correspondence (Yates 1982). We assessed the calibration of predicted probabilities in three ways: 1) analysis of the calibration slope of the predicted risk (Cox 1958; Steyerberg et al. 2001), 2) calibration curve visualization, 3) calculation of the index of prediction accuracy (IPA), a transformation of the Brier score (Kattan et al. 2018). We did not conduct a Hosmer-Lemeshow group-based calibration test due to its recognized limitations and recommendations against its use (Kramer et al. 2007; Van Calster et al. 2019).

Clinical utility

The planned clinical use of the prediction model was first to assess eligibility for early discharge among low-risk patients. Accurately estimating the risk of MACE for patients would allow those low-risk patients to be discharged and avoid additional unnecessary workup, freeing up resources (clinical attention, testing capacity, etc.) for higher risk patients. Low

risk was generally defined as being below a 0.5% well-calibrated probability of MACE within 60 days, with less conservative thresholds of 1% and 2% as additional options.

Our model needed to balance two trade-offs: 1) false "negatives" in which a patient was identified as low-risk but whose true risk of MACE within 60 days was above the threshold, and 2) false "positives" in which patients were believed to be above the given threshold but whose true risk was less than the threshold. Errors in the first category have a greater cost than those in the second category, because there is a greater potential detriment to those patients who were discharged early but whose true risk exceeded the threshold. Patients incorrectly estimated to be above the risk threshold, but who are truly low risk, have comparatively minor costs of additional workup, use of clinical resources, and potential to be overtreated. Yet these possible errors are not quite the same as false negative or false positives typically used to assess predictive models: we care about the true, but unknown, risk rather than the observed outcome. Under this decision-making calculus a patient whose true risk is correctly predicted to be below the clinical threshold, and is therefore discharged without additional workup, but who ends up having a MACE would still have been managed appropriately.

This suggests that the absolute or squared error of the patient's predicted risk versus true risk, particularly near the clinical threshold, would be reasonable loss functions to translate into clinical utility. Miscalibration near the clinical threshold needs to be avoided, whereas miscalibration away from the threshold does not affect the decision. As the expected value of that loss approaches zero we would see that the number of false negatives and false positives (in terms of risk above or below a threshold rather than the observed outcome) also approaches zero. We could target a specific threshold by focusing on patients on the incorrect side of the threshold and averaging the error in their risk prediction, possibly including differential weights for each side of the threshold to account for different costs to the patient. Such a "miscalibration-around-a-threshold" loss function might look as follows:

$$\begin{aligned}\text{loss}(Y_i, \hat{Y}_i \mid X_i) =& \omega_1 \mathbf{1}(P_0(Y_i \mid X_i) < \tau) g\Big(\hat{f}(X_i) - \tau\Big) + \\ & \omega_2 \mathbf{1}(P_0(Y_i \mid X_i) > \tau) g\Big(\tau - \hat{f}(X_i)\Big)\end{aligned} \tag{4.1}$$

where:

- Y is the observed outcome and X is the set of predictors,
- i indexes each patient in the sample,
- $P_0(Y_i \mid X_i)$ is the true risk of patient i,
- $\hat{f}(X_i)$ is the predicted risk of patient i from a given estimator $\hat{f}$,
- τ is the clinical threshold (e.g. 0.5%),

- g is a function such as the identity, squared value, or absolute value function,
- ω_1 is the differential cost for low-risk patients who are kept for further workup,
- ω_2 is the differential cost for high-risk patients who are incorrectly discharged early,

An alternative form of $g(\cdot)$ might operate on the ratio of predicted versus true risk, rather than the absolute difference.

We do not know the true risk for any patients, but we can estimate it within our sample by fitting a semi-parametric smooth function (e.g. lowess) to estimate the true probability of the outcome given the estimated predicted probability, equivalent to what is done during calibration analysis.

If multiple decisions were to be made based on the estimated risk, we might sum this loss over each decision. Alternatively we might use a threshold-free loss function, such as:

$$\text{loss}(Y_i, \hat{Y}_i \mid X_i) = g\Big(\hat{f}(X_i) - \tau\Big) \tag{4.2}$$

Interpretability

Beyond the statistical performance of a clinical prediction, it can be important to provide an explanation or overview of how a model generates its predictions. Interpretation is desirable first because it can provide evidence that the model is working as expected, which can improve the trustworthiness of its predictions for clinicians, patients, or collaborators. Interpretation may also lead to scientific insights about how predictors are related to the outcome, which could be conceptualized as causal pathways, data generating processes, or biological mechanisms. Interpretation can further inform the data export and cleaning processes, such as identifying extreme values, data entry errors, or outliers, or suggesting additional predictor variables to incorporate into the model.

Methods of interpretation can be model-specific or model-agnostic. For models within the family of linear regression, one might provide the estimated beta coefficients for each predictor, along with their associated confidence intervals and p-values. Interpretation becomes less straightforward as models become more complex, such as with interaction or polynomial terms in a regression, random forest or boosted tree models with hundreds or thousands of non-linear decision trees, or splines in which ranges of a given predictor might have different coefficients. The tension is that while restricted model forms, like those assumed by some parametric regression, might make interpretability easier on the assumption that it contains the *true model*, the interpretability becomes moot if the model is misspecified. What are desired are variable importance measures that are not tied to the algorithm fitting the data.

In this work we focus on two complementary forms of model interpretability: *variable importance ranking* and *accumulated local effect plots*, as described below.

Variable importance ranking

Prediction-oriented variable importance rankings order the predictor variables by their contribution to a model's prediction, providing evidence as to which predictors were relied upon the most by the algorithm. Such rankings could be used as a form of confirmatory analysis if a hypothesized ranking were created prior to data analysis, which could formally identify predictors that differed from their expected importance.

Accumulated local effects

It may also be helpful to understand how a model's prediction varies over the values of individual predictors, particularly continuous predictors with a wide range or large number of unique values. Partial dependence plots (PDPs) as proposed by Friedman (2001) are commonly used to provide this type of interpretability, but they can yield flawed results because they make a key unrealistic assumption that features are statistically independent of each other (Molnar 2020, p. 5.1.3). Accumulated local effect plots are a recently developed method that avoids that limitation of PDPs, by counterfactually modifying observations that lie within a nearby kernel neighborhood of the current predictor's value of interest (Apley et al. 2019).

4.3 Results

Model performance

Discrimination

Figure 4.1 displays the estimated precision-recall area under the curve (PR-AUC) PR-AUCs and 95% confidence intervals for each combination of features and estimation algorithm. The MACE mean on the training sample represents the baseline PR-AUC, which was 1.88%. The SuperLearner ensemble achieved the highest estimated PR-AUC (0.148, 95% CI [0.126, 0.170]), followed by the random forest with hyperparameter tuning (0.144, [0.125, 0.164]), the default random forest (0.143, [0.122, 0.165]), and the tuned XGBoost (0.138, [0.116, 0.160]). By comparison the PR-AUC for logistic regression was 0.120 [0.103, 0.137], somewhat lower than the ensemble. Point estimates and confidence intervals are listed in Supplemental Table 4.5.

Figure 4.1. Comparison of cross-validated discriminative performance using PR-AUC metric, with 95% confidence intervals.

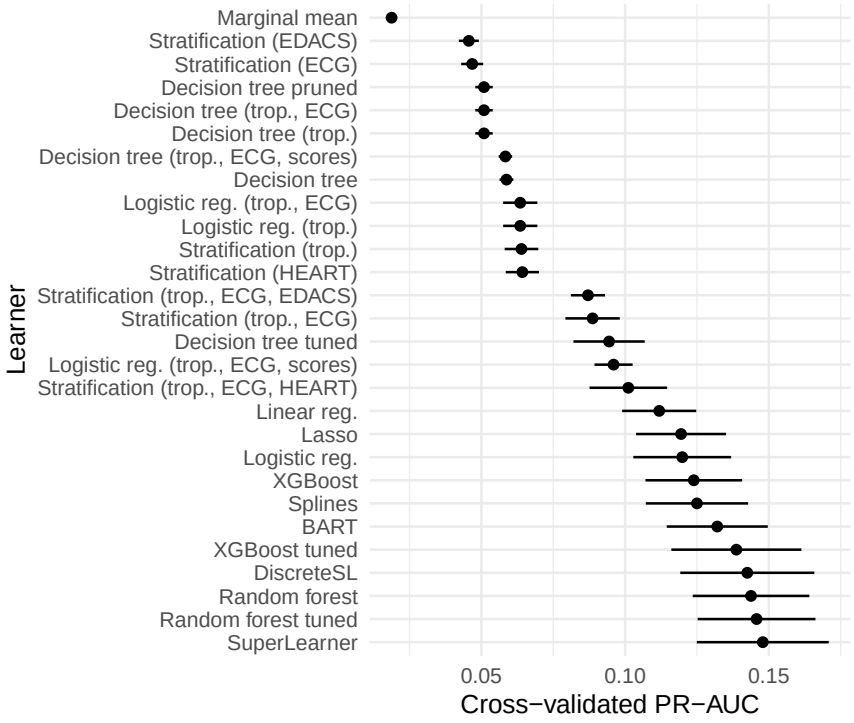
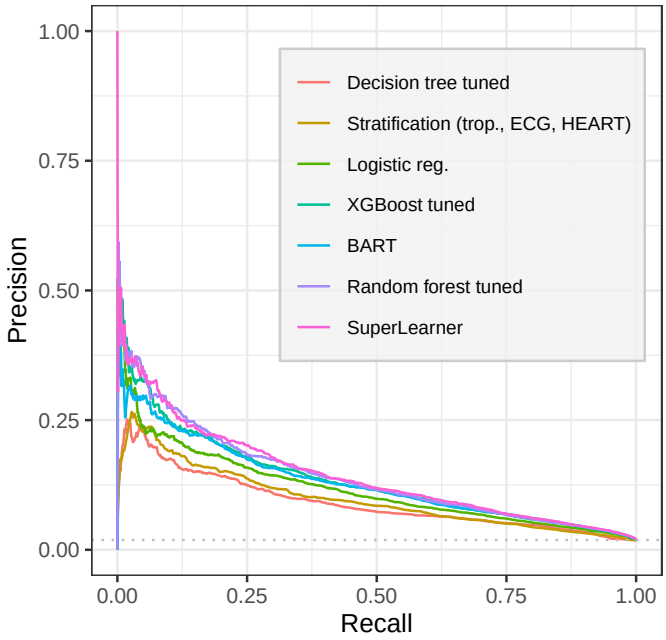
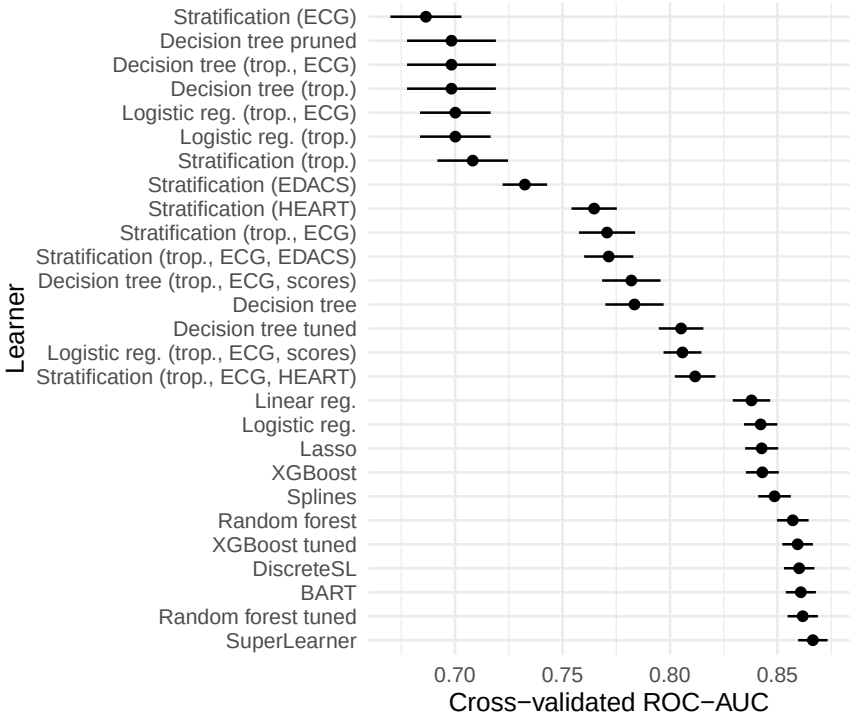


Figure 4.2. Comparison of cross-validated discriminative performance using precision-recall curves for a subset of learners.



For our secondary discrimination metric, cross-validated ROC-AUC was calculated and is displayed in Figure 4.3 (LeDell et al. 2015). The SuperLearner ensemble again achieved the highest performance (ROC-AUC = 0.866, 95% CI [0.859, 0.873]), followed by tuned random forest (0.860, [0.853, 0.867]), tuned XGBoost (0.859, [0.852, 0.866]), and BART (0.859, [0.852, 0.866]). The ROC-AUC for logistic regression (0.842, [0.834, 0.850]) was notably lower than the ensemble. Point estimates and confidence intervals are listed in Supplemental Table 4.6.

Figure 4.3. Comparison of cross-validated discriminative performance using ROC-AUC metric, with 95% confidence intervals. The simple mean had a standard AUC of 0.5 and is omitted from the plot.



We reviewed the distribution of learner weights in the SuperLearner ensemble to examine which algorithms were used most heavily. The weight distribution of learners that were included at least once (i.e. maximum weight greater than 0) is reported in Table 4.1. Four algorithms were always incorporated into the ensemble: default random forest (average weight = 0.25), tuned random forest (average weight = 0.25), default XGBoost (average weight = 0.20), and BART (average weight = 0.18). The remaining 3 learners were sometimes incorporated into the ensemble, with average weights ranging from 0.08 to < 0.01 and a maximum individual weight of 0.18.

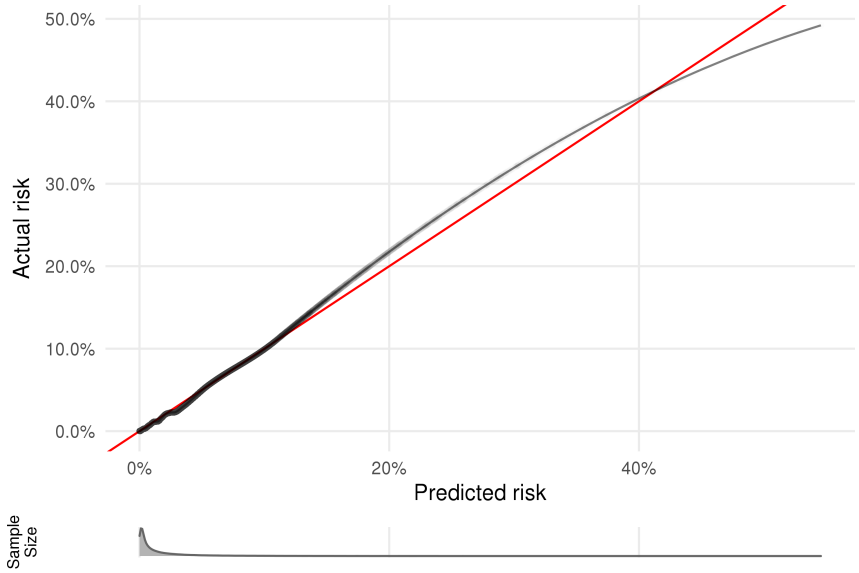
Calibration

We visually compared the predicted risk of the SuperLearner ensemble to the lowess-smoothed observed risk in figure 4.4. The red line is the target calibration in which predicted risk is equal to observed risk. The blue line shows the lowess-smoothed observed risk for each value of the predicted risk.

Table 4.1. Distribution of algorithm weights across ensemble cross-validation replications

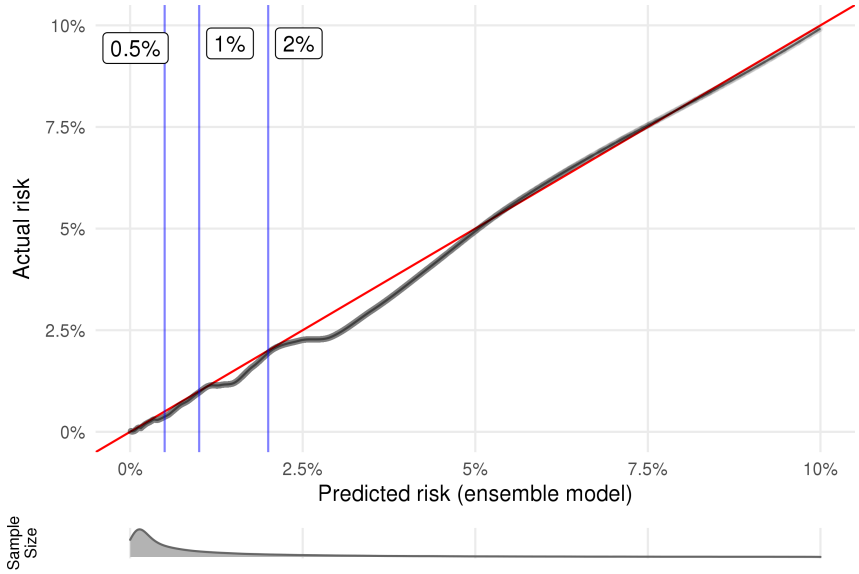
#	Learner	Mean	SD	Min	Max
1	Random forest	0.2516	0.0765	0.1815	0.3749
2	Random forest tuned	0.2488	0.0828	0.1449	0.3737
3	XGBoost	0.1959	0.0378	0.1506	0.2378
4	BART	0.1824	0.0535	0.0993	0.2316
5	Splines	0.0748	0.0685	0.0000	0.1764
6	Logistic reg.	0.0429	0.0309	0.0000	0.0720
7	Stratification (trop., ECG, EDACS)	0.0036	0.0080	0.0000	0.0178

Figure 4.4. Calibration plot comparing predicted risk to observed risk



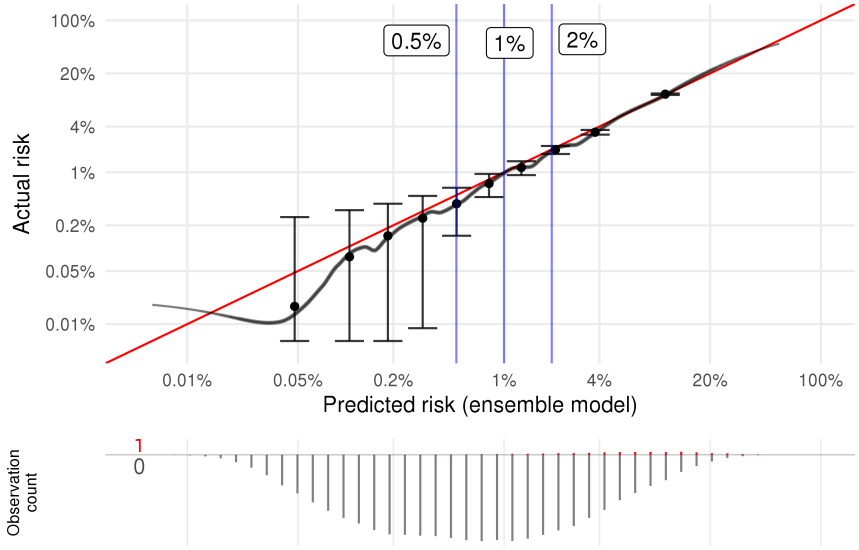
The median predicted risk was 0.64%, with a first quartile of 0.2% and third quartile of 2%. Our primary threshold of scientific interest was 0.5% for possible early discharge. Given those low risk levels, it would be best to "zoom in" our visual calibration review to that region. We show a zoomed calibration plot as Figure 4.5.

Figure 4.5. Zoomed calibration plot comparing predicted risk to observed risk. Clinical thresholds of 0.5%, 1%, or 2% risk are noted by blue vertical lines.



Finally, we include a exponential-scale calibration plot (Figure 4.6) with calibration confidence intervals after grouping patients into 10 groups based on predicted risk, consistent with TRIPOD guideline recommendations (Collins et al. 2015). Due to the substantial class imbalance the exponential scaling of axes allows easier comparison across the probability range, although it may be less intuitive due to the shifting of scales. For example, the width of confidence intervals is counterintuitive for visual comparison due to the dynamic scaling, but the amount of information provided is visually consistent throughout the plot.

Figure 4.6. Exponential-scale calibration plot comparing predicted risk to observed risk with grouped 95% confidence intervals. Clinical thresholds of 0.5%, 1%, or 2% risk are noted by blue vertical lines.



As a statistical complement to the visual examination, we also calculated mean absolute error (MAE). MAE is the sample mean of the absolute difference between the smoothed observed risk (Risk_0) and the predicted risk (Risk_P).

$$\frac{1}{n}\sum_{i=1}^{n} \mid \text{Risk}_O(i) - \text{Risk}_P(i) \mid \tag{4.3}$$

We found an MAE of 0.19% with a lowess smoothing span of 0.05 (low smoothing), and an MAE of 0.14% with a smoothing span of 0.20 (high smoothing). These statistics indicate that the ensemble risk prediction was typically miscalibrated by about 0.17 percentage points.

Missing data imputation

We evaluated the benefit of the more complex GLRM-based imputation by comparing the imputed value to the known value, among variables with missingness [CK: this should be uprgaded to use cross-validation, so this current analysis is speculative]. The root mean-squared error metric was calculated for each variable, and for both GLRM and median/mode imputation methods. We could then estimate the percentage improvement in RMSE for the GLRM imputation. Results in Table 4.2 show a notable improvement in RMSE for every variable, with the exception of the obesity binary variable.

Table 4.2. Comparing missing value imputation using GLRM versus median/mode

Variable	Missingness	Error GLRM	Error Median	Percent reduction
HbA1c	59.8	0.088	1.628	94.6
Triglycerides	46.9	0.039	0.528	92.6
HDL	43.5	1.077	14.815	92.7
Total Chol.	42.7	7.961	45.762	82.6
LDL	38.8	4.158	37.291	88.8
GFR	8.8	0.286	11.699	97.6
Trop. 3HV	2.7	0.001	0.008	90.2
ECG	1.8	0.000	0.729	100.0
BMI	1.6	0.645	7.508	91.4
Obese	1.4	0.769	0.640	-20.1
Respiration	0.4	0.020	2.909	99.3
O2 Saturation	0.3	0.020	2.515	99.2
Pulse	0.2	0.680	18.446	96.3
Pulse Peak	0.2	0.703	18.553	96.2
SBP	0.1	0.586	22.843	97.4
Lowest SBP	0.1	0.458	18.673	97.5

Interpretation

Variable importance ranking

As discussed earlier, our objective for the variable importance analysis was to understand which variables were most influential on the prediction of our final model. Providing that ranking could improve the interpretability of the risk prediction, allowing for confirmation that the results are reasonable and possibly yielding additional scientific insights. However, our final model is quite complex: it is a weighted average of multiple versions of random forests, xgboost models, bayesian additive regression trees, etc. In this work we provide rankings for the top two estimation algorithms: random forest and xgboost. We used the optimal hyperparameter settings from cross-validated analysis.

Interestingly we see rather different results between the two algorithms, which supports the hypothesis than an ensemble of multiple algorithms could achieve better performance than selecting a single estimation algorithm. Both algorithms place high emphasis on the EDACS and HEART risk scores, demonstrating the benefit of including those scores along with the underlying predictors. Different versions of the cardiac troponin predictor are emphasized by the two algorithms: random forest focuses on 3-hour troponin whereas xgboost focuses on peak troponin. ECG reading is emphasized by xgboost but not random forest. Both algorithms make use of lipid profile predictors (LDL, HDL) and vital signs (pulse, respiration) that are not included in the existing risk scores. The random forest makes use

Table 4.3. Variable importance rankings

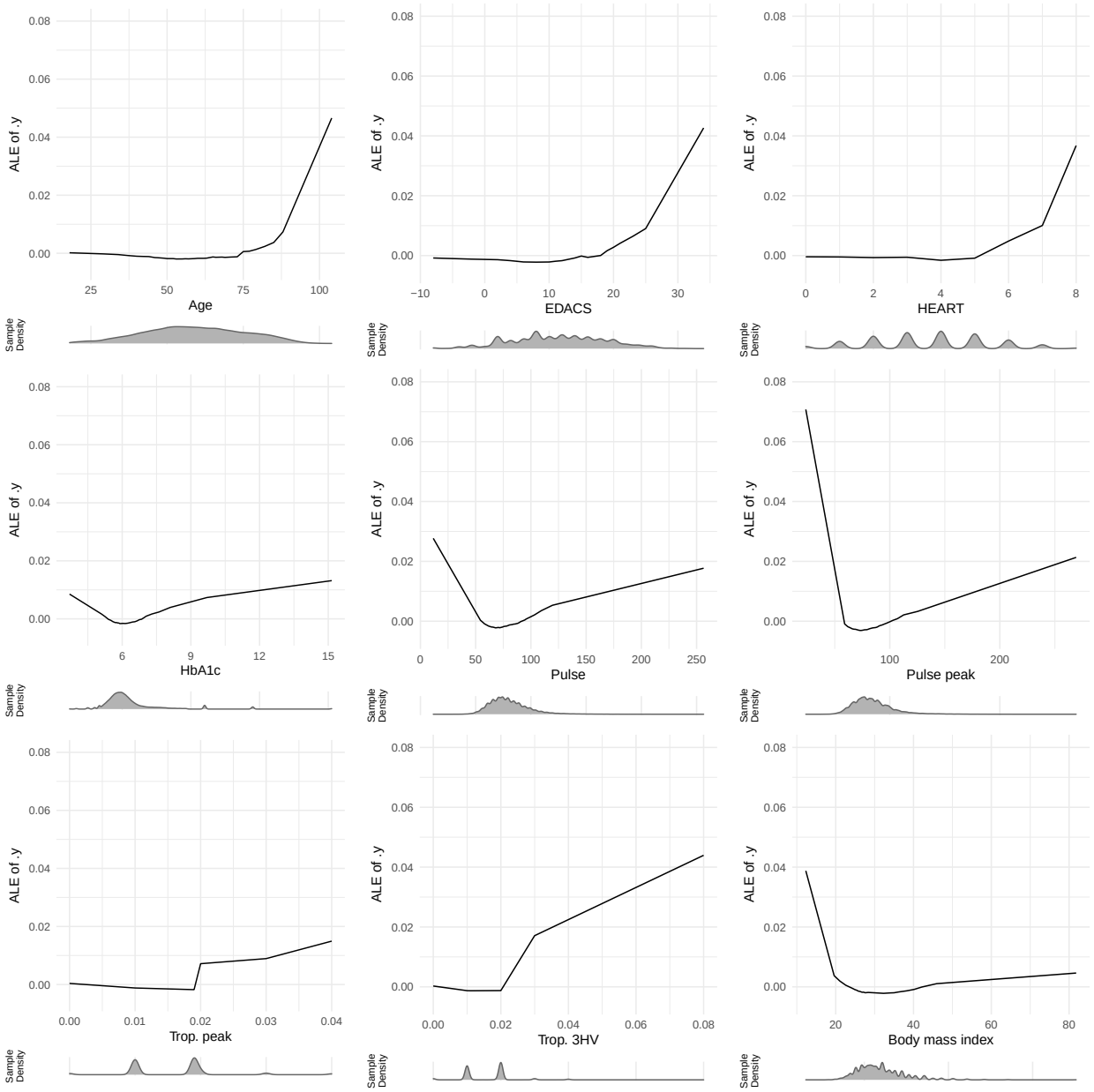
Random Forest importance ranking		XGBoost importance ranking	
Variable	Mean Decrease Accuracy (%)	Variable	Gain
1. Age	0.262	1. Peak troponin	0.3288
2. EDACS	0.188	2. HEART	0.1591
3. HEART	0.117	3. High EDACS	0.0712
4. CAD	0.117	4. EDACS	0.0457
5. Troponin 3HV	0.111	5. High HEART	0.0444
6. LDL	0.104	6. ECG	0.0415
7. Total Cholesterol	0.102	7. Peak pulse	0.0320
8. Missing Triglcyerides	0.083	8. Age	0.0282
9. Missing Total Cholesterol	0.080	9. BMI	0.0234
10. Missing LDL	0.076	10. SBP	0.0217
11. Missing HDL	0.073	11. Myocardial infarction	0.0210
12. Pulse	0.071	12. CAD	0.0198
13. Peak pulse	0.070	13. Aortic athero.	0.0175
14. BMI	0.051	14. Troponin 3HV	0.0165
15. Diabetes	0.047	15. HDL	0.0159
16. Hypertension	0.045	16. Respiration	0.0135
17. HDL	0.044	17. O2 saturation	0.0131
18. HbA1c	0.043	18. GFR	0.0130
19. Peak troponin	0.041	19. Exertion	0.0127
20. Triglycerides	0.039	20. Lowest SBP	0.0091

of certain missingness indicators, which are often indicative of the quality of a patient's records, while xgboost does not. Also noteworthy is the lack of pain-related characteristics sourced from clinical notes in the top predictors, a difference from prior work highlighting their importance at predicting MACE (Amsterdam et al. 2010).

Accumulated local effects

Below we visualize the conditional relationship of top predictors to the ensemble's prediction, across their range of values.

Figure 4.7. Accumulated local effect plots of key continuous predictors



4.4 Discussion

The next step in model evaluation is to conduct one or more external validations of the discrimination and calibration of the model predictions. This validation might include future retrospective cohorts at the current study location (temporal validation), although preferably cohorts sourced from other regions or EHRs (geographical or institutional validation) (Moons et al. 2012). We hope to collaborate in the future with groups interested in such validations.

In future work we plan to expand the machine learning in several ways. The ensemble weighting could specifically optimize PR-AUC. Incorporating feature selection may benefit the simpler algorithms by removing unhelpful predictors. Feature engineering might be beneficial as well, such as creation of interaction terms or even incorporation of the principal components from the GLRM imputation. Due to computational limitations we were not able to conduct hyperparameter tuning on the BART learner, which likely would provide some performance benefit. We are optimistic that random search or model-based search (e.g. Bayesian optimization) rather than grid search could provide even stronger tuning of algorithm hyperparameters across a higher number of dimensions. Evaluation of the GLRM imputation could be further contextualized through comparisons to additional imputation methods, especially principal component analysis, k-nearest neighbors,multiple imputation, and variable-specific supervised models (e.g. OLS or random forest). Additional machine learning algorithms could be explored, such as LightGBM, extremely randomized trees, and multivariate adaptive regression splines. The variable importance ranking could be streamlined through a Random Forest-style permutation importance analysis of the SuperLearner ensemble itself, or through a targeted learning method such as vimp (Williamson et al. 2017) or varimpact (Hubbard, C. J. Kennedy, et al. 2018).

The model might also benefit from a broader sample that includes higher risk patients, which were not included in this study. Calibration might be improved through targeted learning-based adjustment (Brooks et al. 2012). Cross-validated estimation of discrimination performance could be improved through cross-validated targeted maximum likelihood estimation (Benkeser, Petersen, et al. 2019).

The cleaning and analysis code for this project has been translated to use a public dataset and is available online at `https://github.com/ck37/Predictive-Modeling-in-R`. Functions to calculate PR-AUC, ROC-AUC, index of prediction accuracy (IPA), and Brier scores for cross-validated SuperLearner ensembles are provided in the open source R package ck37r, available at `http://github.com/ck37/ck37r`.

4.5 Appendix

Predictor summary

This table summarizes the predictors used in the model following missing value imputation and histogram condensing of unique values for high-cardinality variables. This table includes 71 total predictors but totals 74 predictors when the race predictor is decomposed from a categorical predictor into 4 indicator predictors.

Table 4.4. Summary of predictors

	Name	Unique values	Mode	Mean	Median	Minimum	Maximum
Biomarkers							
1	ECG	3	0	0	0	0	2
2	Peak troponin	6	0	0	0	0	0
3	Troponin 3HV	29	0	0	0	0	0
4	Troponin categories	2	1	1	1	1	2
Scores							
5	EDACS	41	8	11	12	-8	34
6	HEART	9	4	4	4	0	8
7	High EDACS	2	0	0	0	0	1
8	High HEART	2	1	1	1	0	1
Labs							
9	GFR	96	70	66	70	2	70
10	HbA1c	83	6	6	6	4	15
11	HDL	148	48	51	50	4	213
12	LDL	137	111	104	103	8	546
13	Total Cholesterol	104	186	181	181	33	1075
14	Triglycerides	86	5	5	5	3	9
Vitals							
15	BMI	100	30	30	28	12	82
16	Lowest SBP	174	120	121	120	42	244
17	O2 saturation	45	100	98	98	16	100
18	Obesity	2	0	0	0	0	1
19	Peak pulse	153	77	84	82	24	269
20	Pulse	168	74	80	78	12	256
21	Respiration	67	18	18	18	2	100
22	SBP	155	145	145	143	56	269
Demographics							
23	Age	87	53	59	59	18	104

Table 4.4. Summary of predictors

	Name	Unique values	Mode	Mean	Median	Minimum	Maximum
24	Male	2	0	0	0	0	1
25	Race	5	White	-	-	-	-
Clinical notes							
26	Diaphoresis	3	0	0	0	0	9
27	Exertion	3	0	0	0	0	9
28	Pain on inspiration	3	0	0	0	0	9
29	Pain on palpation	3	0	0	0	0	9
30	Radiating pain	6	0	1	0	0	9
31	Sharp pain	3	0	0	0	0	9
History							
32	Anxiety	2	0	0	0	0	1
33	Aortic athero.	2	0	0	0	0	1
34	APD	2	0	0	0	0	1
35	CAD	2	0	0	0	0	1
36	Coronary rev.	2	0	0	0	0	1
37	Diabetes	2	0	0	0	0	1
38	Family history of CAD	2	0	0	0	0	1
39	Hypercholesteremia	2	1	1	1	0	1
40	Hypertension	2	1	1	1	0	1
41	Myocardial infarction	2	0	0	0	0	1
42	PAD	2	0	0	0	0	1
43	Pre-diabetes	2	0	0	0	0	1
44	Prior catheter	2	0	0	0	0	1
45	Prior CT	2	0	0	0	0	1
46	Prior echo	2	0	0	0	0	1
47	Prior MPI	2	0	0	0	0	1
48	Prior treadmill	2	0	0	0	0	1
49	Prior UTL	2	0	0	0	0	1
50	Smoking	2	0	0	0	0	1
51	Stroke	2	0	0	0	0	1
Missingness indicators							
52	Missing BMI	2	0	0	0	0	1
53	Missing Diaphoresis	2	0	0	0	0	1
54	Missing ECG	2	0	0	0	0	1
55	Missing Exertion	2	1	1	1	0	1
56	Missing GFR	2	0	0	0	0	1
57	Missing HbA1c	2	1	1	1	0	1

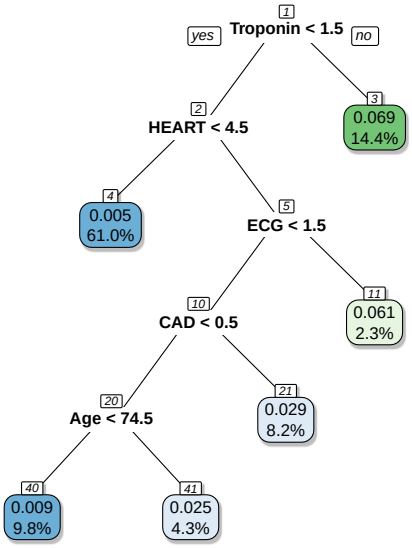
Table 4.4. Summary of predictors

	Name	Unique values	Mode	Mean	Median	Minimum	Maximum
58	Missing HDL	2	0	0	0	0	1
59	Missing LDL	2	0	0	0	0	1
60	Missing O2 Saturation	2	0	0	0	0	1
61	Missing Obesity	2	0	0	0	0	1
62	Missing Pain on Inspiration	2	1	1	1	0	1
63	Missing Pain on Palpation	2	1	1	1	0	1
64	Missing Pulse	2	0	0	0	0	1
65	Missing Radiating Pain	2	0	0	0	0	1
66	Missing Respiration	2	0	0	0	0	1
67	Missing SBP	2	0	0	0	0	1
68	Missing Sharp Pain	2	1	1	1	0	1
69	Missing Total Cholesterol	2	0	0	0	0	1
70	Missing Triglcyerides	2	0	0	0	0	1
71	Missing Troponin 3HV	2	0	0	0	0	1

Decision tree benchmark

A single decision tree is a helpful benchmark approach to risk prediction. It is easy to explain and to visualize, and represents a decision-making process that can be conducted by medical staff. It is limited by its simplicity though: decision trees rarely have competitive predictive performance and the selected variables & cutpoints are known to be sensitive to small changes to the dataset. Figure 4.8 shows an example tree as estimated on our data.

Figure 4.8. Single decision tree, where each leaf node shows 1) predicted risk, and 2) the percentage of patients that fall into that node. Here Troponin refers to peak Troponin.

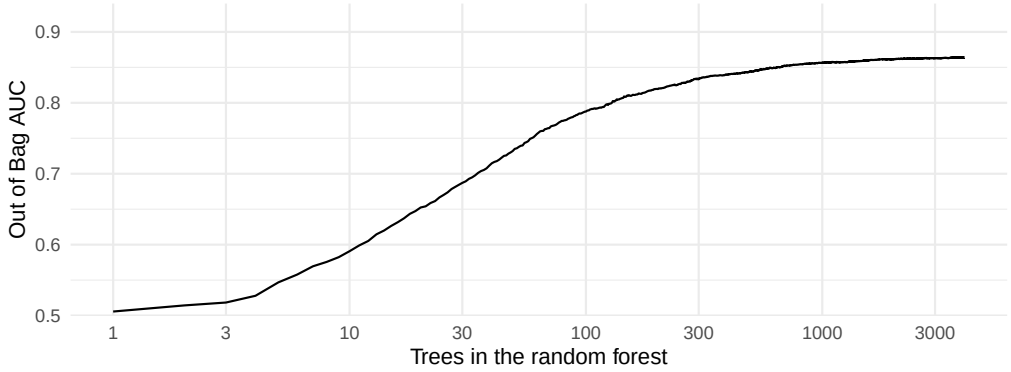


Of particular clinical interest is the blue leaf node (#4) showing a predicted risk of 0.5% and encompassing 61% of the sample. That node represents the lowest-risk group as identified by the decision tree. On first glance the identification of the low-risk group as comprising 61% of patients would seem to improve upon the 50% low-risk result from our complex nested ensemble. However, the 0.5% estimated risk for that leaf node is an average across the large proportion of patients that met the two simple decisions: non-elevated troponin and HEART score ≤ 4. We know that every such patient doesn't have exactly 0.5% risk - there is a distribution around that average, with some patients have lower risk and others having higher risk. Thus we can see that this low-risk node must suffer from some miscalibration: an unknown proportion of those patients have higher risk than our target 0.5% threshold, and should not be discharged early despite the recommendation of the decision tree.

Random forest convergence plot

The number of trees is a key hyperparameter for random forest. Breiman (2001) proved that adding additional trees does not lead to overfitting due to the independence of the tree estimators, unlike the sequential fitting of boosting approaches. Yet it remains beneficial to examine the convergence of the random forest's performance to ensure that an adequate number of trees is being used Probst and Boulesteix (2017). To assess performance convergence we analyzed the out-of-bag ROC-AUC for up to 3,000 trees, shown in Figure 4.9. From visual inspection we can see that only 100 trees would be inadequate, whereas 1,000 - 3,000 trees achieves an ROC-AUC above 0.85.

Figure 4.9. Random forest convergence plot shows ROC-AUC on out-of-bag data as forest size is increased exponentially.



Precision-Recall AUC table

Table 4.5. Cross-validated PR-AUC discrimination performance

Learner	PR-AUC	Std. Err.	CI Lower	CI Upper
Marginal mean	0.0188	0.0000	0.0187	0.0188
Stratification (EDACS)	0.0456	0.0018	0.0422	0.0491
Stratification (ECG)	0.0468	0.0019	0.0430	0.0506
Decision tree (trop.)	0.0509	0.0015	0.0479	0.0539
Decision tree (trop., ECG)	0.0509	0.0015	0.0479	0.0539
Decision tree pruned	0.0509	0.0015	0.0479	0.0539
Decision tree (trop., ECG, scores)	0.0583	0.0012	0.0560	0.0607
Decision tree	0.0587	0.0012	0.0563	0.0611
Logistic reg. (trop.)	0.0635	0.0030	0.0576	0.0695
Logistic reg. (trop., ECG)	0.0635	0.0030	0.0576	0.0695
Stratification (trop.)	0.0639	0.0030	0.0581	0.0698
Stratification (HEART)	0.0643	0.0030	0.0585	0.0700
Stratification (trop., ECG, EDACS)	0.0871	0.0030	0.0812	0.0930
Stratification (trop., ECG)	0.0887	0.0048	0.0792	0.0982
Decision tree tuned	0.0944	0.0063	0.0820	0.1068
Logistic reg. (trop., ECG, scores)	0.0960	0.0034	0.0894	0.1026
Stratification (trop., ECG, HEART)	0.1011	0.0069	0.0876	0.1146
Linear reg.	0.1119	0.0066	0.0990	0.1248
Lasso	0.1195	0.0080	0.1038	0.1351
Logistic reg.	0.1199	0.0087	0.1029	0.1369
XGBoost	0.1239	0.0086	0.1071	0.1407
Splines	0.1250	0.0091	0.1072	0.1428
BART	0.1321	0.0090	0.1145	0.1496
XGBoost tuned	0.1387	0.0115	0.1161	0.1613
Random forest	0.1438	0.0103	0.1235	0.1641
Random forest tuned	0.1457	0.0105	0.1252	0.1662
SuperLearner	0.1479	0.0117	0.1250	0.1709

Receiver Operating Characteristic AUC table

Index of prediction accuracy table

Table 4.6. Cross-validated ROC-AUC discrimination performance

Learner	ROC-AUC	Std. Err.	CI Lower	CI Upper
Marginal mean	0.5000	0.0108	0.4789	0.5211
Stratification (ECG)	0.6864	0.0084	0.6699	0.7029
Decision tree (trop.)	0.6983	0.0105	0.6777	0.7190
Decision tree (trop., ECG)	0.6983	0.0105	0.6777	0.7190
Decision tree pruned	0.6983	0.0105	0.6777	0.7190
Logistic reg. (trop.)	0.7001	0.0084	0.6837	0.7166
Logistic reg. (trop., ECG)	0.7001	0.0084	0.6837	0.7166
Stratification (trop.)	0.7082	0.0084	0.6918	0.7246
Stratification (EDACS)	0.7325	0.0053	0.7221	0.7428
Stratification (HEART)	0.7647	0.0054	0.7541	0.7753
Stratification (trop., ECG)	0.7707	0.0067	0.7577	0.7838
Stratification (trop., ECG, EDACS)	0.7715	0.0058	0.7601	0.7829
Decision tree (trop., ECG, scores)	0.7820	0.0069	0.7685	0.7956
Decision tree	0.7835	0.0069	0.7699	0.7971
Decision tree tuned	0.8052	0.0053	0.7948	0.8156
Logistic reg. (trop., ECG, scores)	0.8058	0.0045	0.7970	0.8146
Stratification (trop., ECG, HEART)	0.8117	0.0048	0.8023	0.8212
Linear reg.	0.8379	0.0044	0.8292	0.8467
Logistic reg.	0.8422	0.0040	0.8344	0.8500
Lasso	0.8427	0.0039	0.8350	0.8504
XGBoost	0.8430	0.0039	0.8353	0.8507
Splines	0.8486	0.0039	0.8411	0.8562
Random forest	0.8571	0.0038	0.8498	0.8645
XGBoost tuned	0.8594	0.0036	0.8523	0.8665
BART	0.8609	0.0036	0.8539	0.8679
Random forest tuned	0.8618	0.0036	0.8547	0.8688
SuperLearner	0.8665	0.0035	0.8596	0.8734

Table 4.7. Cross-validated index of prediction accuracy for each learner and the ensemble

Learner	IPA	Std. Err.	CI Lower	CI Upper
Marginal mean	0.0000	0.0000	0.0000	0.0000
Stratification (EDACS)	0.0138	0.0010	0.0118	0.0158
Stratification (ECG)	0.0153	0.0015	0.0125	0.0182
Logistic reg. (trop.)	0.0189	0.0020	0.0149	0.0229
Logistic reg. (trop., ECG)	0.0189	0.0020	0.0149	0.0229
Decision tree (trop.)	0.0234	0.0016	0.0203	0.0266
Decision tree (trop., ECG)	0.0234	0.0016	0.0203	0.0266
Decision tree pruned	0.0234	0.0016	0.0203	0.0266
Stratification (HEART)	0.0243	0.0017	0.0210	0.0277
Stratification (trop., ECG, EDACS)	0.0253	0.0038	0.0179	0.0327
Stratification (trop.)	0.0273	0.0022	0.0230	0.0317
Decision tree (trop., ECG, scores)	0.0278	0.0011	0.0256	0.0301
Decision tree	0.0283	0.0010	0.0263	0.0302
Stratification (trop., ECG)	0.0384	0.0029	0.0327	0.0440
Decision tree tuned	0.0393	0.0024	0.0345	0.0441
Logistic reg. (trop., ECG, scores)	0.0408	0.0018	0.0371	0.0444
XGBoost	0.0416	0.0092	0.0237	0.0596
Stratification (trop., ECG, HEART)	0.0433	0.0041	0.0352	0.0514
Linear reg.	0.0458	0.0018	0.0423	0.0492
Logistic reg.	0.0539	0.0047	0.0447	0.0631
Lasso	0.0544	0.0041	0.0463	0.0625
Splines	0.0565	0.0054	0.0459	0.0672
BART	0.0641	0.0047	0.0548	0.0733
XGBoost tuned	0.0660	0.0062	0.0539	0.0782
DiscreteSL	0.0661	0.0040	0.0582	0.0740
Random forest	0.0673	0.0035	0.0603	0.0742
Random forest tuned	0.0682	0.0029	0.0626	0.0738
SuperLearner	0.0725	0.0045	0.0637	0.0813

Brier score table

Table 4.8. Cross-validated Brier score for each learner and the ensemble

Learner	Brier score	Std. Err.	CI Lower	CI Upper
Marginal mean	0.01840	0.00001	0.01839	0.01842
Stratification (EDACS)	0.01815	0.00001	0.01812	0.01818
Stratification (ECG)	0.01812	0.00003	0.01807	0.01817
Logistic reg. (trop.)	0.01806	0.00004	0.01797	0.01814
Logistic reg. (trop., ECG)	0.01806	0.00004	0.01797	0.01814
Decision tree (trop.)	0.01797	0.00003	0.01791	0.01804
Decision tree (trop., ECG)	0.01797	0.00003	0.01791	0.01804
Decision tree pruned	0.01797	0.00003	0.01791	0.01804
Stratification (HEART)	0.01796	0.00003	0.01790	0.01801
Stratification (trop., ECG, EDACS)	0.01794	0.00006	0.01781	0.01807
Stratification (trop.)	0.01790	0.00004	0.01781	0.01799
Decision tree (trop., ECG, scores)	0.01789	0.00002	0.01785	0.01793
Decision tree	0.01788	0.00002	0.01785	0.01792
Stratification (trop., ECG)	0.01770	0.00006	0.01759	0.01781
Decision tree tuned	0.01768	0.00005	0.01759	0.01777
Logistic reg. (trop., ECG, scores)	0.01765	0.00003	0.01759	0.01772
XGBoost	0.01764	0.00017	0.01730	0.01798
Stratification (trop., ECG, HEART)	0.01761	0.00008	0.01745	0.01776
Linear reg.	0.01756	0.00003	0.01750	0.01763
Logistic reg.	0.01741	0.00009	0.01724	0.01758
Lasso	0.01740	0.00008	0.01725	0.01755
Splines	0.01736	0.00010	0.01717	0.01756
BART	0.01722	0.00009	0.01705	0.01740
XGBoost tuned	0.01719	0.00012	0.01696	0.01742
DiscreteSL	0.01719	0.00008	0.01703	0.01734
Random forest	0.01717	0.00007	0.01703	0.01730
Random forest tuned	0.01715	0.00006	0.01704	0.01726
SuperLearner	0.01707	0.00009	0.01690	0.01724

Grouped calibration table

Decile	Predicted risk	Observed risk	Predicted / Actual	Predicted - Actual
1	0.0005	0.0002	2.7813	0.0003
2	0.0011	0.0008	1.3707	0.0003
3	0.0018	0.0015	1.2685	0.0004
4	0.0031	0.0025	1.2329	0.0006
5	0.0050	0.0039	1.3012	0.0012
6	0.0080	0.0071	1.1301	0.0009
7	0.0129	0.0116	1.1115	0.0013
8	0.0211	0.0198	1.0660	0.0013
9	0.0376	0.0337	1.1159	0.0039
10	0.1041	0.1067	0.9759	-0.0026

Bibliography

Fisher, Ronald Aylmer (1934). "Two new properties of mathematical likelihood". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 144.852, pp. 285–307.

Brier, Glenn W (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Horvitz, Daniel G and Donovan J Thompson (1952). "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260, pp. 663–685.

Rasch, G (1953). "On simultaneous factor analysis in several populations". In: *Uppsala Symposium on Psychological Factor Analysis, 17-19 March 1953*. Taylor & Francis, pp. 65–71.

Cox, David R (1958). "Two further applications of a model for binary regression". In: *Biometrika* 45.3/4, pp. 562–565.

Sanders, Frederick (1963). "On subjective probability forecasting". In: *Journal of Applied Meteorology* 2.2, pp. 191–201.

Forgy, Edward W (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". In: *Biometrics* 21, pp. 768–769.

Murphy, Allan H and Robert L Winkler (1977). "Reliability of subjective probability forecasts of precipitation and temperature". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 26.1, pp. 41–47.

McCullagh, Peter (1980). "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2, pp. 109–127.

Rasch, G (1980). "Probabilistic models for some intelligence and attainment tests. 1960". In: *Copenhagen, Denmark: Danish Institute for Educational Research*.

Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D Phillips (1981). *Calibration of probabilities: The state of the art to 1980*. Tech. rep. Decision Research. Eugene, OR.

Masters, Geoff N (1982). "A Rasch model for partial credit scoring". In: *Psychometrika* 47.2, pp. 149–174.

Wright, Benjamin D and Geoff N Masters (1982). *Rating scale analysis*. MESA press. URL: `https://research.acer.edu.au/measurement/2/`.

Yates, J Frank (1982). "External correspondence: Decompositions of the mean probability score". In: *Organizational Behavior and Human Performance* 30.1, pp. 132–156.

Cohen, Jacob (1983). "The cost of dichotomization". In: *Applied psychological measurement* 7.3, pp. 249–253.

Linacre, John M (1987). "An extension of the Rasch model to multi-facet situations". In: *Chicago: University of Chicago, Department of Education*.

— (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press. URL: `https://www.winsteps.com/a/Linacre-MFRM-book.pdf`.

Hastie, Trevor J and Robert J Tibshirani (1990). *Generalized additive models*. Vol. 43. CRC press.

Friedman, Jerome H et al. (1991). "Multivariate adaptive regression splines". In: *The annals of statistics* 19.1, pp. 1–67.

Becher, Heiko (1992). "The concept of residual confounding in regression models and some applications". In: *Statistics in medicine* 11.13, pp. 1747–1758.

Muraki, Eiji (1992). "A generalized partial credit model: Application of an EM algorithm". In: *ETS Research Report Series* 1992.1, pp. i–30.

Wang, R, J Sedransk, and JH Jinn (1992). "Secondary data analysis when there are missing observations". In: *Journal of the American Statistical Association* 87.420, pp. 952–961.

Wolpert, David H (1992). "Stacked generalization". In: *Neural networks* 5.2, pp. 241–259.

Wright, Benjamin D (1992). "The International Objective Measurement Workshops: Past and future". In: *Objective measurement: Theory into practice*. Ed. by Mark Wilson. Vol. 1. Norwood, NJ: Ablex Publishing Corporation. Chap. 2.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Breiman, Leo (1996). "Stacked regressions". In: *Machine learning* 24.1, pp. 49–64.

Henning, Grant (1996). "Accounting for nonsystematic error in performance ratings". In: *Language Testing* 13.1, pp. 53–61.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Wright, Benjamin D (1996). "Comparing Rasch measurement and factor analysis". In: *Structural Equation Modeling: A Multidisciplinary Journal* 3.1, pp. 3–24.

Caruana, Rich (1997). "Multitask learning". In: *Machine learning* 28.1, pp. 41–75.

Martin, JA, Glenn Regehr, Richard Reznick, Helen Macrae, John Murnaghan, Carol Hutchison, and M Brown (1997). "Objective structured assessment of technical skill (OSATS) for surgical residents". In: *British journal of surgery* 84.2, pp. 273–278.

Greenland, Sander, Judea Pearl, James M Robins, et al. (1999). "Causal diagrams for epidemiologic research". In: *Epidemiology* 10, pp. 37–48.

Linacre, John M (1999). "Investigating rating scale category utility". In: *Journal of outcome measurement* 3, pp. 103–122.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

Steyerberg, Ewout W, Frank E Harrell Jr, Gerard JJM Borsboom, MJC Eijkemans, Yvonne Vergouwe, and J Dik F Habbema (2001). "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis". In: *Journal of clinical epidemiology* 54.8, pp. 774–781.

Wilson, Mark and Machteld Hoskens (2001). "The rater bundle model". In: *Journal of Educational and Behavioral Statistics* 26.3, pp. 283–306.

Linacre, John M (2002). "What do infit and outfit, mean-square and standardized mean?" In: *Rasch Measurement Transactions* 16.2, p. 878. URL: `https://www.rasch.org/rmt/rmt162f.htm`.

Linacre, John M and Benjamin D Wright (2002). "Construction of measures from many-facet data." In: *Journal of Applied Measurement*.

Streiner, David L (2002). "Breaking up is hard to do: the heartbreak of dichotomizing continuous data". In: *The Canadian Journal of Psychiatry* 47.3, pp. 262–266.

Tsesis, Alexander (2002). *Destructive messages: How hate speech paves the way for harmful social movements*. Vol. 778. NYU Press.

Binderup, Mona-Lise, M Dalgaard, LO Dragsted, A Hossaini, Ole Ladefoged, Henrik Rye Lam, JC Larsen, C Madsen, Otto A Meyer, ES Rasmussen, et al. (2003). "Combined actions and interactions of Chemicals in Mixtures: the toxicological effects of exposure to mixtures of industrial and environmental chemicals". In:

Myford, Carol M and Edward W Wolfe (2003). "Detecting and measuring rater effects using many-facet Rasch measurement: Part I". In: *Journal of applied measurement* 4.4, pp. 386–422.

Senn, Stephen (2003). "Disappointing dichotomies". In: *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 2.4, pp. 239–240.

van der Laan, Mark J and Katherine S Pollard (2003). "A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap". In: *Journal of Statistical Planning and Inference* 117.2, pp. 275–303.

Wood, Simon N (2003). "Thin plate regression splines". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 95–114.

Chang, Hasok (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Hirano, Keisuke and Guido W Imbens (2004). "The propensity score with continuous treatments". In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164, pp. 73–84.

Wilson, Mark (2004). *Constructing measures: An item response modeling approach*. Routledge.

Senn, Stephen (2005). "Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials". In: *Proceedings of the International Statistical Institute, 55th Session, Sydney*.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

Royston, Patrick, Douglas G Altman, and Willi Sauerbrei (2006). "Dichotomizing continuous predictors in multiple regression: a bad idea". In: *Statistics in medicine* 25.1, pp. 127–141.

van der Laan, Mark J and Daniel Rubin (2006). "Targeted maximum likelihood learning". In: *The international journal of biostatistics* 2.1.

Austin, Peter C (2007). "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality". In: *Statistics in medicine* 26.15, pp. 2937–2957.

Cook, Nancy R (2007). "Use and misuse of the receiver operating characteristic curve in risk prediction". In: *Circulation* 115.7, pp. 928–935.

Kortenkamp, Andreas (2007). "Ten years of mixing cocktails: a review of combination effects of endocrine-disrupting chemicals". In: *Environmental health perspectives* 115.Suppl 1, pp. 98–105.

Kramer, Andrew A and Jack E Zimmerman (2007). "Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited". In: *Critical care medicine* 35.9, pp. 2052–2056.

van der Laan, Mark J, Eric C Polley, and Alan E Hubbard (2007). "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1.

Costa, Joaquim F Pinto da, Hugo Alonso, and Jaime S Cardoso (2008). "The unimodal model for the classification of ordinal data". In: *Neural Networks* 21.1, pp. 78–91.

Hubbard, Alan E and Mark J van der Laan (2008). "Population intervention models in causal inference". In: *Biometrika* 95.1, pp. 35–47.

Pan, Sinno Jialin and Qiang Yang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Steyerberg, Ewout W (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media.

Agresti, Alan (2010). *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons.

Amsterdam, Ezra A, J Douglas Kirk, David A Bluemke, Deborah Diercks, Michael E Farkouh, J Lee Garvey, Michael C Kontos, James McCord, Todd D Miller, Anthony Morise, et al. (2010). "Testing of low-risk patients presenting to the emergency department with chest pain: a scientific statement from the American Heart Association". In: *Circulation* 122.17, pp. 1756–1776.

Chipman, Hugh A, Edward I George, Robert E McCulloch, et al. (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Rozenholc, Yves, Thoralf Mildenberger, and Ursula Gather (2010). "Combining regular and irregular histograms by penalized likelihood". In: *Computational Statistics & Data Analysis* 54.12, pp. 3313–3323.

Zheng, Wenjing and Mark J van der Laan (2010). "Asymptotic theory for cross-validated targeted maximum likelihood estimation". In:

Andrich, David (2011). "Rating scales and Rasch measurement". In: *Expert review of pharmacoeconomics & outcomes research* 11.5, pp. 571–585.

Goos, Peter and Bradley Jones (2011). *Optimal design of experiments: a case study approach*. John Wiley & Sons.

Segal, Mark and Yuanyuan Xiao (2011). "Multivariate random forests". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 80–87.

van der Laan, Mark J and Sherri Rose (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization". In: *Journal of machine learning research* 13.Feb, pp. 281–305.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz (2012). "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk". In: *Political analysis* 20.3, pp. 351–368.

Brooks, Jordan, Mark J van der Laan, and Alan S Go (2012). "Targeted maximum likelihood estimation for prediction calibration". In: *The international journal of biostatistics* 8.1.

Dawson, Neal V and Robert Weiss (2012). "Dichotomizing continuous variables in statistical analysis: a practice to avoid." In:

Moons, Karel GM, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward (2012). "Risk prediction models: II. External validation, model updating, and impact assessment". In: *Heart* 98.9, pp. 691–698.

Muñoz, Iván Díaz and Mark van der Laan (2012). "Population intervention causal effects based on stochastic interventions". In: *Biometrics* 68.2, pp. 541–549.

Warner, William and Julia Hirschberg (2012). "Detecting hate speech on the world wide web". In: *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, pp. 19–26.

Bien, Jacob, Jonathan Taylor, and Robert Tibshirani (2013). "A lasso for hierarchical interactions". In: *Annals of statistics* 41.3, p. 1111.

Campello, Ricardo JGB, Davoud Moulavi, and Jörg Sander (2013). "Density-based clustering based on hierarchical density estimates". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 160–172.

Carlin, Danielle J, Cynthia V Rider, Rick Woychik, and Linda S Birnbaum (2013). *Unraveling the health effects of environmental mixtures: an NIEHS priority*.

Engelhard Jr, George (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Kennedy, Edward H, Wyndy L Wiitala, Rodney A Hayward, and Jeremy B Sussman (2013). "Improved cardiovascular risk prediction using nonparametric regression and electronic health record data". In: *Medical care* 51.3, p. 251.

Metayer, Catherine, Joanne S Colt, Patricia A Buffler, Helen D Reed, Steve Selvin, Vonda Crouse, and Mary H Ward (2013). "Exposure to herbicides in house dust and risk of childhood acute lymphoblastic leukemia". In: *Journal of Exposure Science and Environmental Epidemiology* 23.4, p. 363.

Scheipl, Fabian, Thomas Kneib, and Ludwig Fahrmeir (2013). "Penalized likelihood and Bayesian function selection in regression models". In: *AStA Advances in Statistical Analysis* 97.4, pp. 349–385.

Stanton, Gregory (2013). *The Ten Stages of Genocide*. Genocide Watch. URL: `https://www.genocidewatch.com/ten-stages-of-genocide`.

Sun, Zhichao, Yebin Tao, Shi Li, Kelly K Ferguson, John D Meeker, Sung Kyun Park, Stuart A Batterman, and Bhramar Mukherjee (2013). "Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons". In: *Environmental Health* 12.1, p. 85.

Than, Martin, Mel Herbert, Dylan Flaws, Louise Cullen, Erik Hess, Judd E Hollander, Deborah Diercks, Michael W Ardagh, Jeffery A Kline, Zea Munro, et al. (2013). "What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: a clinical survey". In: *International journal of cardiology* 166.3, pp. 752–754.

Whitehead, Todd P, F Reber Brown, Catherine Metayer, June-Soo Park, Monique Does, Myrto X Petreas, Patricia A Buffler, and Stephen M Rappaport (2013). "Polybrominated diphenyl ethers in residential dust: sources of variability". In: *Environment international* 57, pp. 11–24.

Bobb, Jennifer F, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull (2014). "Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures". In: *Biostatistics* 16.3, pp. 493–508.

Henn, Birgit Claus, Brent A Coull, and Robert O Wright (2014). "Chemical mixtures and childrens health". In: *Current opinion in pediatrics* 26.2, p. 223.

Hilden, Jørgen and Thomas A Gerds (2014). "A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index". In: *Statistics in medicine* 33.19, pp. 3405–3414.

Kerr, Kathleen F, Zheyu Wang, Holly Janes, Robyn L McClelland, Bruce M Psaty, and Margaret S Pepe (2014). "Net reclassification indices for evaluating risk-prediction instruments: a critical review". In: *Epidemiology (Cambridge, Mass.)* 25.1, p. 114.

Lawson, John (2014). *Design and Analysis of Experiments with R*. Chapman and Hall/CRC.

Leening, Maarten JG, Moniek M Vedder, Jacqueline CM Witteman, Michael J Pencina, and Ewout W Steyerberg (2014). "Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide". In: *Annals of internal medicine* 160.2, pp. 122–131.

Than, Martin, Dylan Flaws, Sharon Sanders, Jenny Doust, Paul Glasziou, Jeffery Kline, Sally Aldous, Richard Troughton, Christopher Reid, William A Parsonage, et al. (2014). "Development and validation of the E mergency D epartment A ssessment of C hest pain S core and 2 h accelerated diagnostic protocol". In: *Emergency Medicine Australasia* 26.1, pp. 34–44.

Bailey, Helen D, Claire Infante-Rivard, Catherine Metayer, Jacqueline Clavel, Tracy Lightfoot, Peter Kaatsch, Eve Roman, Corrado Magnani, Logan G Spector, Eleni Th. Petridou, et al. (2015). "Home pesticide exposures and risk of childhood leukemia: Findings from the childhood leukemia international consortium". In: *International journal of cancer* 137.11, pp. 2644–2663.

Carrico, Caroline, Chris Gennings, David C Wheeler, and Pam Factor-Litvak (2015). "Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting". In: *Journal of agricultural, biological, and environmental statistics* 20.1, pp. 100–120.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang (2015). "Xgboost: extreme gradient boosting". In: *R package version 0.4-2*, pp. 1–4.

Collins, Gary S, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons (2015). "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement". In: *British Journal of Surgery* 102.3, pp. 148–158.

Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati (2015). "Hate speech detection with comment embeddings". In: *Proceedings of the 24th international conference on world wide web*. ACM, pp. 29–30.

Eckes, Thomas (2015). *Introduction to many-facet Rasch measurement,* 2nd Ed. Frankfurt: Peter Lang.

LeDell, Erin, Maya Petersen, and Mark van der Laan (2015). "Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates". In: *Electronic journal of statistics* 9.1, p. 1583.

Liu, Bing (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Pepe, Margaret S, Jing Fan, Ziding Feng, Thomas Gerds, and Jorgen Hilden (2015). "The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets". In: *Statistics in biosciences* 7.2, pp. 282–295.

Saito, Takaya and Marc Rehmsmeier (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3.

Benkeser, David and Mark van der Laan (2016). "The highly adaptive lasso estimator". In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, pp. 689–696.

Braun, Joseph M, Chris Gennings, Russ Hauser, and Thomas F Webster (2016). "What can epidemiological studies tell us about the impact of chemical mixtures on human health?" In: *Environmental health perspectives* 124.1, A6–A9.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Fong, Christian and Justin Grimmer (2016). "Discovery of treatments from text corpora". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1600–1609.

Goldstein, Benjamin A, Ann Marie Navar, and Rickey E. Carter (July 2016). "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges". In: *European Heart Journal* 38.23, pp. 1805–1814. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehw302.

Goldstein, Benjamin A, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis (May 2016). "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association* 24.1, pp. 198–208. ISSN: 1067-5027. DOI: 10.1093/jamia/ocw042.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Hou, Le, Chen-Ping Yu, and Dimitris Samaras (2016). "Squared earth mover's distance-based loss for training deep neural networks". In: *arXiv preprint arXiv:1611.05916*.

Hubbard, Alan E, Sara Kherad-Pajouh, and Mark J van der Laan (2016). "Statistical inference for data adaptive target parameters". In: *The international journal of biostatistics* 12.1, pp. 3–19.

Hubbard, Alan E and Mark J van der Laan (2016). *Mining with Inference: Data-Adaptive Target Parameters.*

Metayer, Catherine, Gary Dahl, Joe Wiemels, and Mark Miller (2016). "Childhood leukemia: a preventable disease". In: *Pediatrics* 138.Supplement 1, S45–S55.

Myers, Raymond H, Douglas C Montgomery, and Christine M Anderson-Cook (2016). *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons.

Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang (2016). "Abusive language detection in online user content". In: *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, pp. 145–153.

Olson, Randal S, Ryan J Urbanowicz, Peter C Andrews, Nicole A Lavender, La Creis Kidd, and Jason H Moore (2016). "Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I". In: ed. by Giovanni Squillero and Paolo Burelli. Springer International Publishing. Chap. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pp. 123–137. ISBN: 978-3-319-31204-0. DOI: 10.1007/978-3-319-31204-0_9. URL: http://dx.doi.org/10.1007/978-3-319-31204-0_9.

Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki (Sept. 2016). "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis". In: *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. Ed. by Michael BeiSSwenger, Michael Wojatzki, and Torsten Zesch. Vol. 17. Bochumer Linguistische Arbeitsberichte. Bochum, pp. 6–9.

Rui, P, K Kang, and JJ. Ashman (2016). *National Hospital Ambulatory Medical Care Survey: 2016 emergency department summary tables*. URL: https://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2016_ed_web_tables.pdf.

Samejima, Fumiko (2016). "Graded response models". In: *Handbook of item response theory, volume one*. Chapman and Hall/CRC, pp. 123–136.

Schuler, Alejandro, Vincent Liu, Joe Wan, Alison Callahan, Madeleine Udell, David E Stark, and Nigam H Shah (2016). "Discovering patient phenotypes using generalized low rank

models". In: *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, pp. 144–155.

Sellars, Andrew (2016). *Defining hate speech*. Tech. rep. Berkman Klein Center. URL: https://dx.doi.org/10.2139/ssrn.2882244.

Taylor, Kyla W, Bonnie R Joubert, Joe M Braun, Caroline Dilworth, Chris Gennings, Russ Hauser, Jerry J Heindel, Cynthia V Rider, Thomas F Webster, and Danielle J Carlin (2016). "Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop". In: *Environmental health perspectives* 124.12, A227–A229.

Udell, Madeleine, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. (2016). "Generalized low rank models". In: *Foundations and Trendső in Machine Learning* 9.1, pp. 1–118.

Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma (2017). "Deep learning for hate speech detection in tweets". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, pp. 759–760.

Beckham, Christopher and Christopher Pal (2017). "Unimodal probability distributions for deep ordinal classification". In: *arXiv preprint arXiv:1705.05278*.

Davalos, Angel D, Thomas J Luben, Amy H Herring, and Jason D Sacks (2017). "Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures". In: *Annals of epidemiology* 27.2, pp. 145–153.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber (2017). "Automated hate speech detection and the problem of offensive language". In: *Eleventh International AAAI Conference on Web and Social Media*.

Del Vigna, Fabio, Andrea Cimino, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi (2017). "Hate me, hate me not: Hate speech detection on Facebook". In:

Hastie, Trevor, Robert Tibshirani, and Ryan J Tibshirani (2017). "Extended comparisons of best subset selection, forward stepwise selection, and the lasso". In: *arXiv preprint arXiv:1707.08692*.

Kennedy, Edward H, Zongming Ma, Matthew D McHugh, and Dylan S Small (2017). "Nonparametric methods for doubly robust estimation of continuous treatment effects". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 1229–1245.

Magu, Rijul, Kshitij Joshi, and Jiebo Luo (2017). "Detecting the hate code on social media". In: *Eleventh International AAAI Conference on Web and Social Media*.

Probst, Philipp and Anne-Laure Boulesteix (2017). "To tune or not to tune the number of trees in random forest". In: *The Journal of Machine Learning Research* 18.1, pp. 6673–6690.

Ranjan, Rajeev, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa (2017). "An all-in-one convolutional neural network for face analysis". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, pp. 17–24.

Ruder, Sebastian (2017). "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Whitehead, Todd P, Praphopphat Adhatamsoontra, Yang Wang, Elisa Arcolin, Leonard Sender, Steve Selvin, and Catherine Metayer (2017). "Home remodeling and risk of childhood leukemia". In: *Annals of epidemiology* 27.2, pp. 140–144.

Williamson, Brian D, Peter B Gilbert, Noah Simon, and Marco Carone (2017). "Nonparametric variable importance assessment using machine learning techniques". In:

Wilson, Richard Ashby (2017). *Incitement on trial: prosecuting international speech crimes*. Cambridge University Press.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Benesch, Susan, Cathy Buerger, Tonei Glavinic, and Sean Manion (2018). "Dangerous Speech: A Practical Guide". In: *Dangerous Speech Project*.

Bobb, Jennifer F, Birgit Claus Henn, Linda Valeri, and Brent A Coull (2018). "Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression". In: *Environmental Health* 17.1, p. 67.

de la Torre, Jordi, Domenec Puig, and Aida Valls (2018). "Weighted kappa loss function for multi-class classification of ordinal data in deep learning". In: *Pattern Recognition Letters* 105, pp. 144–154.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman (2018). "Measuring and mitigating unintended bias in text classification". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, pp. 67–73.

Engelhard Jr, George and Stefanie Wind (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Fortuna, Paula and Sérgio Nunes (2018). "A survey on automatic detection of hate speech in text". In: *ACM Computing Surveys (CSUR)* 51.4, p. 85.

Founta, Antigoni-Maria, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis (2018). "A unified deep learning architecture for abuse detection". In: *arXiv preprint arXiv:1802.00385*.

Greenslade, Jaimi H, Edward W Carlton, Christopher Van Hise, Elizabeth Cho, Tracey Hawkins, William A Parsonage, Jillian Tate, Jacobus Ungerer, and Louise Cullen (2018). "Diagnostic accuracy of a new high-sensitivity troponin I assay and five accelerated diagnostic pathways for ruling out acute myocardial infarction and acute coronary syndrome". In: *Annals of emergency medicine* 71.4, pp. 439–451.

Howard, Jeremy and Sebastian Ruder (2018). "Universal language model fine-tuning for text classification". In: *arXiv preprint arXiv:1801.06146*.

Hubbard, Alan E, Chris J Kennedy, and Mark J van der Laan (2018). "Data-Adaptive Target Parameters". In: *Targeted Learning in Data Science*. Springer, pp. 125–142.

Johnson, Kipp W, Jessica Torres Soto, Benjamin S Glicksberg, Khader Shameer, Riccardo Miotto, Mohsin Ali, Euan Ashley, and Joel T Dudley (2018). "Artificial intelligence in cardiology". In: *Journal of the American College of Cardiology* 71.23, pp. 2668–2679.

Kattan, Michael W and Thomas A Gerds (2018). "The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models". In: *Diagnostic and prognostic research* 2.1, p. 7.

Mark, Dustin G, Jie Huang, Uli Chettipally, Mamata V Kene, Megan L Anderson, Erik P Hess, Dustin W Ballard, David R Vinson, and Mary E Reed (2018). "Performance of coronary risk scores among patients with chest pain in the emergency department". In: *Journal of the American College of Cardiology* 71.6, pp. 606–616.

Molnar, Christoph (2018). "Interpretable machine learning". In: *A Guide for Making Black Box Models Explainable*.

Palomaki, Jennimaria, Olivia Rhinehart, and Michael Tseng (2018). "A Case for a Range of Acceptable Annotations." In: *SAD/CrowdBias@ HCOMP*, pp. 19–31.

Pearl, Judea and Dana Mackenzie (2018). *The book of why: the new science of cause and effect*. Basic Books.

Ruder, Sebastian (July 2018). *NLP's ImageNet moment has arrived*. URL: `https://thegradient.pub/nlp-imagenet/`.

van der Laan, Mark J and David Benkeser (2018). "Highly adaptive lasso (hal)". In: *Targeted Learning in Data Science*. Springer, pp. 77–94.

van der Laan, Mark J, Aurélien Bibaut, and Alexander R Luedtke (2018). "CV-TMLE for Nonpathwise Differentiable Target Parameters". In: *Targeted Learning in Data Science*. Springer, pp. 455–481.

van der Laan, Mark J and Sherri Rose (2018). *Targeted learning in data science*. Springer.

Weisskopf, Marc G, Ryan M Seals, and Thomas F Webster (2018). "Bias amplification in epidemiologic analysis of exposure to mixtures". In: *Environmental health perspectives* 126.4, p. 047003.

Zhang, Ziqi, David Robinson, and Jonathan Tepper (2018). "Detecting hate speech on twitter using a convolution-gru based deep neural network". In: *European Semantic Web Conference*. Springer, pp. 745–760.

Apley, Daniel W and Jingyu Zhu (2019). "Visualizing the effects of predictor variables in black box supervised learning models". In: *arXiv preprint arXiv:1612.08468*.

Aroyo, Lora, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen (2019). "Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions". In: *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, pp. 1100–1105.

Benkeser, David, Maya Petersen, and Mark J van der Laan (2019). "Improved small-sample estimation of nonlinear cross-validated prediction metrics". In: *Journal of the American Statistical Association*, pp. 1–16.

Blencowe, NS, L Rooshenas, Z Tolkien, KD Bera, H Gould Brown, D Elliott, BC Reeves, and JM Blazeby (2019). "A qualitative study to identify indicators of the quality of wound closure". In: *Journal of Infection Prevention* 20.5, pp. 214–223.

Christodouloua, Evangelia, MA Jie, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, Ben van Calster, et al. (2019). "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". In: *Journal of clinical epidemiology*.

Geva, Mor, Yoav Goldberg, and Jonathan Berant (2019). "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets". In: *arXiv preprint arXiv:1908.07898*.

He, Jianxing, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang (2019). "The practical implementation of artificial intelligence technologies in medicine". In: *Nature medicine* 25.1, p. 30.

Heinzerling, Benjamin (2019). *NLP's Clever Hans Moment has Arrived*. The Gradient. URL: https://thegradient.pub/nlps-clever-hans-moment-has-arrived/.

Karpathy, Andrej (2019). "Multi-Task Learning in the Wilderness". ICML. URL: https://slideslive.com/38917690/multitask-learning-in-the-wilderness.

Lazarevic, Nina, Adrian G Barnett, Peter D Sly, and Luke D Knibbs (2019). "Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: a review of existing approaches and new alternatives". In: *Environmental health perspectives* 127.2, p. 026001.

Lazarevic, Nina, Luke D Knibbs, Peter D Sly, and Adrian G Barnett (2019). "Performance of variable and function selection methods for estimating the non-linear health effects of correlated chemical mixtures: a simulation study". In: *arXiv preprint arXiv:1908.01583*.

Linacre, John M (2019). *Facets computer program for many-facet Rasch measurement, version 3.83.0*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Molnar, Christoph (2019). *Interpretable machine learning*. Lulu.com.

Niven, Timothy and Hung-Yu Kao (July 2019). "Probing Neural Network Comprehension of Natural Language Arguments". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4658–4664. DOI: 10.18653/v1/P19-1459. URL: https://www.aclweb.org/anthology/P19-1459.

Olson, Randal S and Jason H Moore (2019). "TPOT: A tree-based pipeline optimization tool for automating machine learning". In: *Automated Machine Learning*. Springer, pp. 151–160.

Perrin, Andrew and Monica Anderson (Apr. 2019). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. Tech. rep. Pew Research Center: Internet, Science & Technology. URL: https://www.pewresearch.org/fact-tank/

2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/.

Polley, Eric C, Erin LeDell, Chris J Kennedy, Sam Lendle, and Mark van der Laan (2019). *Package SuperLearner*.

Probst, Philipp, Anne-Laure Boulesteix, and Bernd Bischl (2019). "Tunability: Importance of Hyperparameters of Machine Learning Algorithms." In: *Journal of Machine Learning Research* 20.53, pp. 1–32.

Probst, Philipp, Marvin N Wright, and Anne-Laure Boulesteix (2019). "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, e1301.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1, p. 8.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *arXiv preprint arXiv:1910.10683*.

Renzetti, Stefano, Chris Gennings, and Paul C Curtin (2019). "gWQS: An R Package for Linear and Generalized Weighted Quantile Sum (WQS) Regression". In: *Journal of Statistical Software*.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith (2019). "The risk of racial bias in hate speech detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678.

Siegel, Alexandra, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker (2019). "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath". In:

Van Calster, Ben, David J McLernon, Maarten van Smeden, Laure Wynants, and Ewout W Steyerberg (2019). "Calibration: the Achilles heel of predictive analytics". In: *BMC medicine* 17.1, pp. 1–7.

Vargas, Victor Manuel, Pedro Antonio Gutierrez, and Cesar Hervas (2019). "Deep Ordinal Classification Based on the Proportional Odds Model". In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, pp. 441–451.

Blake, Helen A., Clémence Leyrat, Kathryn E. Mansfield, Laurie A. Tomlinson, James Carpenter, and Elizabeth J. Williamson (2020). "Estimating treatment effects with partially observed covariates using outcome regression with missing indicators". In: *Biometrical Journal*. DOI: 10.1002/bimj.201900041.

Janssens, A Cecile J W and Forike K Martens (Jan. 2020). "Reflection on modern methods: Revisiting the area under the ROC Curve". In: *International Journal of Epidemiology*. dyz274. ISSN: 0300-5771. DOI: 10.1093/ije/dyz274.

Keil, Alexander P, Jessie P Buckley, Katie M OBrien, Kelly K Ferguson, Shanshan Zhao, and Alexandra J White (2020). "A quantile-based g-computation approach to addressing the effects of exposure mixtures". In: *Environmental health perspectives* 128.4, p. 047004.

Klyuchnikov, Nikita, Ilya Trofimov, Ekaterina Artemova, Mikhail Salnikov, Maxim Fedorov, and Evgeny Burnaev (2020). "Nas-bench-nlp: Neural architecture search benchmark for natural language processing". In: *arXiv preprint arXiv:2006.07116*.

Molnar, Christoph (2020). *Interpretable Machine Learning*. Lulu. com.

Sun, Lichao, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong (2020). "Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT". In: *arXiv preprint arXiv:2003.04985*.

Westling, Ted (2020). "Nonparametric tests of the causal null with non-discrete exposures". In: *arXiv preprint arXiv:2001.05344*.