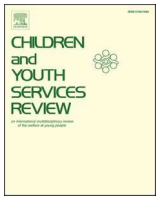
Children and Youth Services Review 158 (2024) 107454

Contents lists available at ScienceDirect

Children and Youth Services Review

journal homepage: www.elsevier.com/locate/childyouth



A computational social science approach to understanding predictors of Chafee service receipt

Jason Yan [a,1], Seventy F. Hall [b], Melanie Sage [b], Yuhao Du [a], Kenneth Joseph [a,*]

[a] Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[b] School of Social Work, University at Buffalo, Buffalo, NY, USA

ARTICLE INFO

Keywords:
Chafee services
Computational social science
Predictive modeling
National Youth in Transition Database

ABSTRACT

The John H. Chafee Foster Care Program for Successful Transition to Adulthood (CFCIP) allocates funding to provide services to youth who are likely to age out of foster care. These services, covering everything from mentoring to financial aid, are expected to be distributed in ways that prepare youth for life after care. One natural question to ask is, *which youth receive Chafee services?* The present work makes use of the National Youth in Transition Database (NYTD), a large-scale administrative dataset that tracks services allocated to youth that use CFCIP funds to answer this question. Specifically, we conduct a *forensic social science* analysis of the NYTD data. To do so, we first use computational methods to help us uncover the factors that best predict which youth will receive services associated with service receipt. We find that the majority of variables in the Adoption and Foster Care Analysis and Reporting System (AFCARS) and NYTD have limited or no utility in predicting Chafee service receipt, and that a subset of three variables—youth age, youth time in care, and the state in which a youth is in care—explain almost all variability in service receipt. We conclude with a discussion of the implications of these and other findings on future research on Chafee service allocation, and the utility of predictive modeling in child welfare, with a particular focus on the utility of the NYTD in this context.

1. Introduction

One of the goals of the U.S. child welfare system is, where appropriate, to keep youth with their families.[2] Despite this goal and efforts to achieve it, tens of thousands of youth each year are removed from their families. Many of these youth will eventually *age out* of foster care without returning to their families or being adopted; that is, they will turn an age at which they are no longer considered wards of the state and eligible for services. Youth who age out of care, on average, have significantly worse life outcomes as compared to general populations of youth. For example, they are known to experience criminal or juvenile justice system involvement, food insecurity, and homelessness at much higher rates than their peers (Lockwood et al., 2015).

Recognizing this, the federal government has passed several laws that aim to improve outcomes among older youth in foster care. The Foster Care Independence Act, signed into law in 1999, established the John H. Chafee Foster Care Program for Successful Transition to Adulthood (CFCIP). Under CFCIP, the federal government allocates Title IV-E funding to states to provide services to foster youth who are likely to age out of care. States in turn use this money to refund foster care agencies, both public and private, for services rendered to foster youth. Services funded by the Chafee program include anything from mentoring to financial assistance for housing and post-secondary education.

A number of studies have explored how services are allocated to foster youth under the CFCIP (Chor et al., 2018; Pérez et al., 2020; Okpych, 2015; Thompson et al., 2021), and what the effects of those services are on youth (Collins and Ward, 2011, 2005, 2021, 2022, 2021, 2019, 2019, 2018). The present work focuses exclusively on the former, asking *what can we say about the youths who do (not) receive Chafee services?* Answering this question is important in order to a) identify patterns in how services are allocated, and b) develop prescriptive recommendations for better or more refined interventions in the future.

One important resource for research in this area is the National Youth in Transition Database (NYTD) (Bureau et al., 2019). Foster care agencies that receive funding from CFCIP are mandated to report data regarding case-level information on the youth and the independent

* Corresponding author.
E-mail address: kjoseph@buffalo.edu (K. Joseph).
[1] Now at the School of Information, University of Michigan, Ann Arbor, MI, USA.
[2] https://www.childwelfare.gov/topics/permanency/reunification/

https://doi.org/10.1016/j.childyouth.2024.107454
Received 23 December 2021; Received in revised form 13 December 2023; Accepted 9 January 2024
Available online 12 January 2024

living services they receive. These services are entered into the NYTD, and coded into one of 15 high-level service types that the database tracks. Data on youth service receipt are collected and recorded bi-annually in the NYTD, giving a set of longitudinal snapshots of allocation patterns over time.

The first and most widely cited (as of writing) use of the NYTD data to study service allocation patterns comes from Okpych (2015). They analyzed Chafee service receipt across the U.S. among all youth ages 16–21 who were in foster care in fiscal year (FY) 2011 and were eligible for Chafee-funded services. Their goal was to identify characteristics of youth who were eligible for the Chafee program who in turn did, or did not, receive services. Using descriptive statistics, their work showed that receipt of specific services varied across race/ethnicity, gender, and whether or not an individual had been identified as having a disability. Their work also showed important effects of the urbanicity of the counties in which youth were in care, and that this effect varied across racial lines.

The present work provides a complementary analysis of NYTD data on Chafee service receipt to that of Okpych (2015). Specifically, we propose a distinct methodological approach that leads to a number of novel findings about the factors that may, and equally as important, may *not* be associated with Chafee service receipt among older youth. Our work points to the potential limitations of using the NYTD and AFCARS to study Chafee service allocation. Specifically, we draw on an analytical paradigm increasingly popular in sociology (e.g. Salganik et al., 2020; Kozlowski et al., 2018), social psychology (e.g. Garg et al., 2018), and political science (e.g. Roberts et al., 2014) that is commonly referred to as *forensic social science* (McFarland et al., 2015). Forensic social science "is an approach that merges applied and theory-driven perspectives… [to] guide deductive explorations of the data while also using induction to discover which theories afford an explanation" (McFarland et al., 2015, pg. 20). Critically, forensic social *science*, as a subfield of computational social science (Lazer et al., 2009; Epstein, 1999), is distinct from forensic social *work*, a separate methodology defined as "the practice specialty in social work that focuses on the law and educating law professionals about social welfare issues and social workers about the legal aspects of their objectives" (Barker, 2003, p. 140), (cf. Roberts and Brownell, 1999).

Where a more traditional social scientific approach, like that pursued by Okpych (2015), uses a small number of independent variables to evaluate a finite set of hypotheses, forensic social science aims to leverage "big data" and modern computational methods to confirm existing claims in new ways and generate new avenues for inquiry and/or practice that were not apparent before the analyses were conducted (Goldberg, 2015). We employ a particular kind of forensic social science analysis here that uses predictive modeling to understand social phenomena (see Salganik et al., 2020, discussed below, for the most elaborate example).

With a predictive modeling approach, we first identify a *prediction task*, in which the goal is to guess (predict) the value of some outcome for each individual in the population of interest, given other information available about the individual. Our interest is in understanding how the factors associated with youth who do (not) receive services translate into the following prediction task: *Given data on foster youth from the NYTD and other associated datasets, can we predict how many distinct services that youth will receive in a given year?* Informed by prior work and empirical observations, we ask this question across three distinct subsets of services.

Having identified a prediction task, we then develop and carry out a *prediction experiment*, in which we evaluate a number of different approaches, or *predictive models*, to making these predictions. These models range in both the complexity and quality of their predictions, from simple baselines (e.g., predicting the same number for all youth), to standard regression models, to *machine learning* models that incorporate dozens of predictors into a complex, non-linear function to make predictions. Our experiment consists of repeated trials, where we separate our population into "training" and "test" sets of individuals. In each trial, we then "teach" the model on the training data (e.g. fit regression models) using the training data, and then estimate how well the models "learn the material" by assessing prediction quality (model performance) on the test data. Intuitively, models that perform better on the test better "understand" the data, and can help us gain insights into it. Finally, we conduct a model exploration where we assess the relative predictive power of the individual-level and structural factors used by the model that best explains the data.

This predictive modeling approach, where we consider computational models that can identify predictive power of factors in the data even when those effects are highly non-linear (e.g. where effects vary by state) can help us to address three complementary questions. First, what are the *most* important factors predicting our outcome, considering these potential non-linearities? To the extent these factors align with and/or are distinct from prior work, a forensic social science approach can help to validate that we are focusing on the correct and complete set of important variables. Second, what are the factors that do not help us predict which youth receive Chafee services, even when a highly complex model is used? Understanding which factors may lack predictive power over our outcome can help us narrow the scope of future inquiry. Finally, forensic social science approaches can help us understand the *limits of predictability within the context of a particular dataset*; in our case, we can understand how well we can make predictions about which youth will receive which services, given data in the NYTD.

Notably, our use of predictive modeling in the present work is distinct from its use in what we will call *point-of-decision predictive modeling*, where the goal is to use predictive modeling to *make or aid in on-the-ground decisions*. Such applications of predictive modeling are common in child welfare. For example, over a dozen child welfare agencies used predictive modeling in 2021 (Samant et al., 2021), with systems in place to help identify youth at risk of abuse (Chouldechova et al., 2018; Vaithianathan et al., 2018), and to help select youth for social services interventions (Saxena et al., 2020a). However, scholars have identified problems with existing tools, including biases in their predictions (Purdy and and Glass, 2020; Rodriguez and Storer, 2020; Yelick and Thyer, 2020; Cheng et al., 2022; Keddell, 2019) and a critical gap between expected use and practice (Kawakami et al., 2022b; Kawakami et al., 2022a). These challenges with point-of-decision predictive modeling have led some to recommend eradicating its use entirely from child welfare (Abdurahman, 2021a). In contrast, the present work argues for a more nuanced perspective, where we use predictive modeling but in a forensic social science setting. Here, we can use predictive models to help us understand the data we have and to provide evidence for or new lines of inquiry, and not by making suggestions at the point in time where decisions are made.

In sum, our work makes three contributions to the literature. First, substantively, we show that variables in the NYTD and associated datasets can be categorized loosely into one of three groups: (1) a small number of factors (three) that are strongly predictive of how many Chafee services youth receive, (2) a modest number of factors that have smaller but non-zero predictive power, and (3) a large number of factors that have almost no observed predictive power. We provide a discussion of the implications of these findings for the study of Chafee service allocation and the utility of predictive modeling as applied to data from the NYTD and associated datasets. Second, methodologically, we establish forensic social science as a principled methodology that can provide new insights into problems in child welfare via the combined use of predictive modeling and domain knowledge. We are careful, however, to compare and contrast this use of predictive modeling from its use at the point of decision-making, emphasizing the distinct considerations in each setting. Finally, we provide open materials[3] that others can use to replicate and extend our work, creating a bridge to new

[3] https://github.com/kennyjoseph/chafee_service_alloc.

explorations that use our methodology.

2. Background

2.1. Prior work on Chafee service allocation with the NYTD

A number of recent efforts have been made to study how Chafee services are allocated using the NYTD. As mentioned previously, Okpych (2015) analyzed Chafee service receipt across the U.S. among youth in foster care (ages 16–21) using data from the 2011–2012 NYTD. They observed biases in service allocation across gender—females were more likely to receive services, race—multi-racial and Hispanic youth were more likely to receive at least one service than any other groups, and Black youth were the least likely to receive a Chafee service— and county urbanicity— youth in rural/non-metropolitan areas were more likely to receive services, including more kinds of services, than youth in large metropolitan areas.

However, using broad descriptive statistics can lead to what is known as *Simpson's paradox*, where patterns in data aggregated across many individuals mask more complex, and sometimes contradictory, patterns at lower levels of aggregation (Lerman, 2018). In part driven by a search for these lower-level patterns, recent work has moved beyond the descriptive efforts of Okpych (2015). In particular, Chor et al. (2018) studied youth in foster care from FY 2011–2013 who received at least one Chafee-funded service according to the NYTD. The authors used a variant of latent class analysis (LCA), multi-level LCA, that accounted for state-level effects to cluster youth based on the set of services they received. Their work identified three service profiles: "High-service receipt", "Limited service receipt", and a class of youth who only received Academic Support or an Independent Living Needs assessment. These classes varied in size, representing around 20%, 30%, and 50% of their data, respectively. Chor et al. (2018) then showed significant predictors of youth falling into each class, finding differences on age, gender, educational attainment, and race. Pérez et al. (2020), using the same methodology as Chor et al. (2018), but with only 16-year-old youth, also identified a three-class clustering of youth by service receipt, with similar underlying factors predicting class membership.

The present work extends prior efforts by including youth who did not receive any services, in addition to those who did, and by using recent iterations of the NYTD dataset. Our analysis considers the importance of nearly all variables in the NYTD and associated datasets in predicting Chafee service receipt. Doing so allows us to provide a more detailed outline of which, out of all possible factors in the available data, most determine Chafee service receipt and guides future explorations of service allocation using these data. We also identify potential limitations to the NYTD and associated datasets in what they may be able to tell us about Chafee service receipt.

2.2. Forensic social science

Forensic social science uses machine learning and social theory to advance understanding of social phenomena. Like Radford and Joseph (2020, pg.1), we define social theory "broadly, as the set of scientifically-defined constructs like race, gender, social class, inequality, family, and institution, and their causes and consequences for one another." We define machine learning in a similarly broad fashion, as any computational model that learns from experience (Mitchell et al., 1997). The field of machine learning therefore encompasses traditional statistical models, like linear regression. However, the focus in machine learning research is distinct from typical uses of linear regression in the social sciences because the focus is typically on *prediction*, rather than *explanation* (Hofman et al., 2021). The goal of machine learning, at least initially, is to find any possible way to predict outcomes instead of explaining how those outcomes were predicted.

This focus on prediction has led to the development of models able to identify relationships between independent and dependent variables that would not have been uncovered with traditional statistical methods (Radford and Joseph, 2020). Similarly, we can be more confident that independent variables not leveraged by these more complex models to predict the outcome are unlikely to play a significant direct role in determining the outcome. Predictive modeling cannot, however, tell us whether a particular variable *causes* or is merely correlated with the outcome and cannot explain why the outcome is occurring. Thus, the findings of work focused on prediction can be misleading when applied to policy and intervention settings (Hofman et al., 2021). In contrast, social theory involves identifying the causal mechanisms that may explain why a particular variable is, or is not, predictive.

Machine learning and social theory can therefore be used productively together, and the field of forensic social science has developed around this intersection. The utility of a forensic social science approach is perhaps best described by McFarland et al. (2015):

> [T]he use of machine learning is atheoretical, but it is potentially powerful when used as an agnostic search for potential explanations. In contrast, theory is a somewhat narrow-minded but powerful tool…[that] affords potential explanations for how features interrelate. As such, the iterative combination of atheoretical induction and theory-led deduction can be quite powerful. (pg. 10).

Perhaps the most well-known empirical forensic social science work is from Salganik et al. (2020). In this work, the authors organized and conducted a competition to predict life outcomes in the Fragile Families dataset, which tracks youth from birth through adulthood. The core finding of this challenge was that despite comparing dozens of complex competing models, none systematically outperformed a simple, baseline model that used a standard logistic regression to predict life outcomes at age 18 from four theoretically-motivated variables: mother's race/ethnicity, marital status, and education level, and the same outcome (or a closely related proxy) measured at age 9. In the present work, we compare our predictive model results to analogous theoretically motivated baseline models, helping us to understand if and how more complex models prompt new insights beyond existing theory. Finding that more complex models *do* perform better than our baseline models, we turn towards an analysis of why we believe this to be the case, and what the practical ramifications of this finding are.

2.3. Predictive modeling in child welfare

As noted, forensic social science uses predictive models to help inform how services are allocated. This differs significantly from the standard use of machine learning in child welfare today, where scholars aim to build machine learning models that can be used at the point of decision. For a general review of this usage, we direct the reader to Saxena et al. (2020b). Here, we provide a brief overview.

Existing efforts to use machine learning in child welfare have primarily attempted to predict the level of risk when a child is referred to a child welfare hotline. Predictions are typically based on historical administrative data from public welfare systems (Teixeira and Boyas, 2017). This use of automated decision-making has elicited significant criticism given the possibility that reliance on past administrative data embeds biases into future predictions of risk (Capatosto, 2017), and reinforces the over-surveillance of families of color and those who use public welfare (Eubanks, 2018; Glaberson, 2019; Abdurahman, 2021a).

Few machine learning tools have been built to guide decision-making at later stages in the child welfare pipeline (e.g., to inform service allocation decisions for older youth). One example is the Think of Us platform recently piloted in Santa Clara County, CA and throughout the state of Nebraska (see Brindley et al., 2018). This tool collects ongoing data from youth on their independent living goals, offers automated recommendations and action steps for youth, permits them to assign supportive adults to assist with individualized plans, and prompts the caseworker to investigate when youth have unmet needs. Despite the dearth of real-world applications of algorithms outside of child welfare

Table 1
A sequential list of inclusion criterion used in this study to identify the final population of interest. The first column in this table lists the criteria used. The second lists how many youth were removed from the sample because of that criteria. The final lists the percentage of the remaining sample removed because of this criteria (i.e. the percent of the sample that resulted from all previous criteria).

Inclusion Criterion	N Removed	% Removed
Younger than 22 at the start of the FY	30	<.001%
Older than 14 at the start of the FY	552,641	80.4%
In care for at least 6 months in FY	49,189	36.5%
First removal date >22 years prior to FY	13	.02%
Removal or Setting Change Dates Not Missing	85	.1%
Race/Ethnicity Data Not Missing	883	1.0%
In State where Services were Recorded (not NC or PR)	2,172	2.6%
Setting Length of Stay Not Missing	931	1.1%
Parents Died Indicator Not Missing	710	0.9%
Total youth removed from study	606,654	88.3%
Total Youth Included: 80,714	**Original Sample Size: 687,368**	

hotline screening, scholars have begun to call for these tools (e.g. Ahn et al., 2021; Andreswari et al., 2018) and federal policy is trending toward embedding machine learning at all levels of the U.S. child welfare system to improve case planning and decision-making (e.g. Harrison et al., 2018). It is thus imperative that we consider potential ramifications of the use of these tools at later stages of the child welfare pipeline, as well.

In child welfare systems and other public benefit bureaucracies, the risks and rewards of using machine learning are particularly felt by service users, who are often already the most vulnerable. These methods therefore deserve additional scrutiny *before* such tools are implemented and can cause harm. Human services administrators often have little training in predictive analytics, and may turn over decision-making to vendors who make lofty promises of efficiency and cost-savings; vendors may deliver black-box machine learning models that under-perform (Kelly, 2017). These methods often fail to fully account for the cultural and organizational contexts in which data are collected and decisions are made (Church and Fairchild, 2017) and/or enforced (Christin, 2017). Significant concerns arise with machine learning models that, while meeting mathematical definitions of fairness, do not necessarily meet practical definitions that account for equity, transparency, or the degree to which non-white youth are over-represented in the child welfare system and suffer adverse outcomes at disproportionate rates (Brown et al., 2019).

These critiques point to a need for caution in using machine learning in the context of child welfare. Here, we are motivated by the potential for machine learning to serve not in an applied decision-making role, i.e. point-of-decision predictive modeling, but rather to better understand and, in turn, theorize, the ways that youth are impacted by decision-making within the existing system, *and what role existing data may (or may not) play in helping us understand this*. However, given the somewhat subtle distinction between the two uses, the present work aims to make explicit how the use of predictive modeling differs between forensic social science and point-of-decision predictive modeling, and what the different considerations may be for each.

3. Data

Our analysis draws on two datasets made available by the National Data Archive on Child Abuse and Neglect (NDACAN). The first is derived from the Adoption and Foster Care Analysis and Reporting System (AFCARS) (Bureau et al., 2018). On a yearly basis, AFCARS provides information about all youth who were in foster care (i.e. any youth that meets the federal definition of 45 CFR 1355.20) that particular year. Second, we use data from the NYTD. The NYTD comprises multiple datasets; here we use data containing a public record of all youth whose receipt of Chafee-funded services were reported to the state. The NYTD *Services data* is released twice per year, and can be linked through an anonymized ID to the AFCARS dataset.

Our analysis focuses primarily on data from the 2018 fiscal year (FY). In this section, we first describe how we select individuals from the full AFCARS dataset for inclusion in our study. We then describe three dependent variables that serve as the outcomes we seek to predict. Finally, we detail the independent variables (or synonymously here, *features*) that we use to make these predictions.

3.1. Inclusion/exclusion criteria

Table 1 lists the set of inclusion criteria that youth were required to meet to be considered for our study. Note that the exclusions are sequential, and percentages in the table reflect the percent of youth removed from the data after all prior steps are considered. The primary cause of exclusion from the study was age; in particular, we analyzed only youth aged 14–22. These youth received the vast majority (over 98%) of all services. The second most important exclusion criterion was being in care for at least 6 months, which ensured all youth in our study had spent enough time in care to have had the opportunity to receive services if they were likely to get them. Finally, we excluded youth based on a number of missing data criteria, where these missing data did not

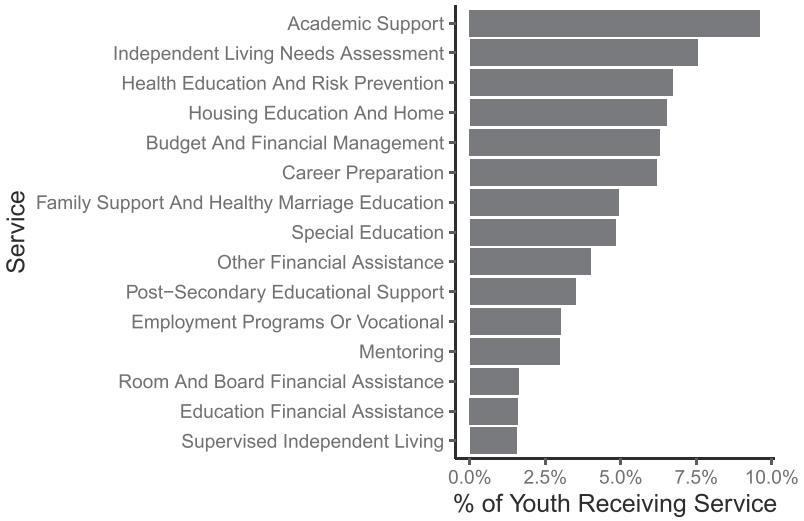


Fig. 1. The percentage of youth (y-axis) listed as having received each Chafee service (y-axis) at least one time in FY 2018.

Table 2
A Summary of all variables used in our models. For this table, we categorized variables into three types: Categorical Variables with meaningful variable levels, binary variables (with an "Unknown" level in some cases), and continuous variables. These three types of variables are shown separately. Categorical variables are listed alongside their various levels; binary and continuous variables are simply listed. Variables and variable levels are separated by semicolons (;).

Categorical Variables

Variable Name	Variable Levels
Race/Ethnicity	NH, White; NH, Am Ind AK Native; NH, Asian; NH, Haw/Pac Isl.; NH, >1 Race; Hispanic
Case Goal	Reunification; Live w/ Relatives; Adoption; Long Term Foster Care; Emancipation; Guardianship; Not Yet Established
Caretaker Family	Married Couple; Unmarried Couple; Single Female; Single Male; Unknown/None
Placement Setting	Pre-adoptive Home; Foster family home, relative; Foster family home, non-relative; Group home; Institution; Supervised independent living; Runaway; Trial home visit
Manner of Removal	Removed Voluntarily; Court Ordered; Unknown Reason for Removal
RU13 Urban/Rural Code	8 Levels, Ranging from 1) Metro: >1 million population to 8) Rural or < 2.5 K population, Non-Adjacent
State	All 50 states, besides North Carolina, Plus DC
Age (Categorical)	14, 15, 16,17, 18, 19+
Age on Date of Legal Adoption	Not Applicable; Less than 2 years old; 2–5 years old; 6–12 years old; 13 years or older; Unable to determine
Discharge Reason	Not Applicable; Reunified with parent, primary caretaker; Living with other relative(s); Adoption; Emancipation; Guardianship; Transfer to another agency; Runaway; Death of child
Sex	Male, Female

Binary Variables (3 Levels for each: Yes/Applies; No/Does not Apply; Unknown)
Sexual Abuse; Physical Abuse; Neglect; Alcohol Abuse Parent; Drug Abuse Parent; Alcohol Abuse Child; Drug Abuse Child; Child Disability; Child Behavioral Problem; Parents Died; Parents in Jail; Caretaker inability; Abandonment; Relinquishment; Inadequate Housing; Parents Rights Terminated; Title IV-E Foster Care Payments; Title IV-E Adoption Assistance; Title IV-A TANF Payment; Title IV-D Child Support Funds; Title XIX Medicaid; SSI or Social Security Act Benefits; Only State or Other Support; Aged Out of Foster Care; Clinical Disability, Mental Retardation; Visual/Hearing Impaired; Physical Disability; Emotionally Disturbed; Other Medical Issue; Current Placement Setting Outside State; Dad's Rights Terminated; Mom's Rights Terminated, In AFCARS Dataset in Previous Year; Was Discharged From Latest Removal; Was Discharged From Previous Removal; Child is Waiting for Adoption; Has Ever Had Periodic Review

Continuous Variables (Both Raw Values and Logged Values)
Previous year service count (all services); Previous year service count (Academic and Employment Support Services); Previous year service count (financial services); Previous year service count (each service individually); Current Setting Length of Stay; Total Number of Removals; Total Number of Placements; First Removal Date; Latest Removal Date; Latest Setting Date; Date of Discharge from Previous Removal; Date of Discharge from Latest Removal; Date of Latest Periodic Review; Age at End of FY

seem reasonable to impute (see below), and excluded youth in North Carolina and Puerto Rico, where Chafee services data were not recorded in FY 2018 in the NYTD.

3.2. Dependent variables

Our prediction experiment explores three different outcome variables drawn from 14 of the 15 services listed in the NYTD Service File. We exclude Special Education services for two reasons. First, special education services are theoretically distinct from other services because they are associated with a school-based assessment instead of a social services assessment. Second, as shown in Appendix A, special education services are empirically distinct in that they are largely uncorrelated with all other services. With the remaining 14 services, we construct three dependent variables:

1. **All Services** - The total number of unique services that a youth receives. The maximum possible value for this outcome is 14 and includes all services in the NYTD dataset, except for Special Education (these are listed on the x-axis of Fig. 1).
2. **Financial Services** - The number of unique services a youth receives from the following set: Supervised Independent Living, Room and Board Financial Assistance, Education Financial Assistance, Other Financial Assistance. These services are unique in that they all either pay a youth directly in cash or pay for a service that the youth would normally pay for themselves to meet everyday needs related to housing and education.
3. **Academic and Employment Support Services** The number of unique services a youth receives from the following set: Academic Support, Post-secondary Educational Support, Career Preparation, Employment Programs or Vocational Training, Budget and Financial Management, Housing Education and Home, Health Education and Risk Prevention, Family Support and Healthy Marriage Education, and Mentoring. These services all provide non-monetary supports, generally in the form of education or social support.

These three outcome variables are interesting for different reasons. All services, together, help us understand factors related to who gets the most and least services. Financial services help us understand services provided directly to youth to practice their own independence, and are often offered to youth who are expected to pay for their own needs, whereas academic and employment support services often seek to promote self-sufficiency through skills-based training programs and are sometimes corrective. For instance, family support and education services are often delivered to youth who are already parents or are likely to become parents. In addition to the theoretical reasons for selecting these subsets of services, we again show in Appendix A that there is empirical support that these services are clustered together in their allocation as well.

3.3. Independent variables

Table 2 lists the 61 categorical, binary, and continuous variables used in our analysis to attempt to predict our dependent variables. Aligning with the forensic social science approach (McFarland et al., 2015), these 61 variables represent as many of the variables in the NDACAN data as possible while maintaining face validity in the context of prior theory and our domain experience. Variables excluded using this criterion consisted exclusively of foster family demographics; all other variables in the data were included. Exclusion of the foster family demographic variables resulted from observations in early modeling work that our predictive models severely overemphasized the importance of this demographic information, leading the models to ignore other variables known to be important, such as placement type. Although we cannot confirm this, our belief is that this was caused by unresolved differences in how states dealt with coding this variable for youth who are not in, or who used to be in, foster care. Nonetheless, removing these foster family variables produced results that better aligned with (while still informing) existing theory, without any significant change in the predictive performance of our models. Thus, we decided to remove them.

Descriptive statistics for each variable are included in the supplementary material provided in the code release for this project. For continuous variables, we include both the raw value and the logarithm of the variable in our predictive models, as is common in predictive analyses where both exact time frames and orders of magnitude for time may provide salient information (Salganik et al., 2020). All continuous variables relevant to dates are measured as days since the end of FY 2018.

Missing values are addressed in one of two ways. Where a reasonable default value could be identified, we replaced missing values with appropriate defaults. For example, 2,380 youth had missing values for

Clinical Disability; these rows were replaced with a value representing "Not Yet Determined" that was also in the AFCARS code book. The full set of imputed variables are provided in the data and code accompanying this article. In a related vein, we do not attempt to impute values for categories marked as "Unknown" or "Not Yet Determined". The primary reason for this is that we do not believe these values to be missing at random, but rather that the missingness of these values is potentially a meaningful signal of how a youth is receiving services within the foster care system (Sankhe et al., 2022). Rather than seek to infer "true" values for these missing quantities, then, we treated missingness itself as a theoretically and empirically meaningful quantity.

4. Methods

We first describe the set of *predictive models* that we construct. Note that in all cases for these predictive models, model parameters are calculated, or "learned" on one subset of the data, the *training data*, and evaluated on another (the *test data*). We then detail how we *evaluate* these models. Finally, we detail our approach to model *exploration*.

4.1. Prediction models

We construct two machine learning models to make predictions about service allocation. Both of our machine learning models are *tree-based*. We opt to focus on tree-based models for two complementary reasons. First, tree-based models are capable of capturing the complex dependencies inherent in our administrative data. An example of this is the variable *Date of Latest Periodic Review*, which is dependent on the variable *Has Ever Had Periodic Review*. In a tree-based model, the algorithm can learn to separate youth into those that have or have not ever had a periodic review, and then make use of information contained in the date of that review. Notably, a linear model (e.g., a linear regression), cannot do this, even with an interaction term, because a coefficient for each independent variable is applied to each youth.

Our first tree-based model is the *Random Forest* (Breiman, 2001). A random forest model is an ensemble of decision trees, where the output of the random forest is the aggregate of all decisions made by the individual decision trees. In the case of a random forest regressor, which we used in our work, the output is simply the average of all the individual decision trees estimate. The idea behind a random forest is that individual decision trees may not be so accurate, but when combined, the output will be closer to the true value on average. Our second tree-based model uses the *Gradient Boosted Trees (GBT)* algorithm (Friedman, 2001), implemented in the computing library `XGBoost`. It is a gradient-boosting algorithm that iteratively combines decision trees as the "weak" predictors to produce a much stronger model. The principal idea is that each decision tree builds upon previous trees by learning the residuals, essentially a correcting term. The final output is the sum of the output from each individual tree.

Both the random forest model and the GBT model we use contain *hyperparameters*, or parameters that are fixed for any one run of the algorithm. There are a number of ways to select hyperparameters. Here, we choose to do so in two different ways to avoid dependency on any one strategy. For random forests, we identify hyperparameters by finding values that allow for the best predictions on services data from prior years. We then fix these hyperparameters for prediction on the 2018 data.[4]. For GBT, instead of optimizing hyperparameters on our data, we opt to select a single setting that has been suggested as a reasonable default by prior work.[5]

In addition to these two models, we construct six simple baseline models derived from reasonable expectations based on prior work described above about 1) how past service receipt should predict future receipt, 2) the impacts of age, race, and rural/urbanicity (as considered by Okpych, 2015), and 3) the impact of the state in which a youth lives on the services they receive (as considered by Chor et al., 2018; Pérez et al., 2020):

1. *Previous Year's Service Count* - For this baseline model, we predict the number of services a youth will receive in 2018 using the number of services they received in 2017. If a youth was not in care in 2017, we predict a value of 0.
2. *Constant* - We first compute the average number of services received by youth in the training data. We then predict that each youth in the test data will receive this amount of services. This model is equivalent to a linear regression model with only an intercept term.
3. *Age* - We first compute the average number of services received by youth in each age category (14, 15, 16, 17, 18, 19+) in the training data. We then use these averages to make predictions in the test data, based on the age category of each youth. This model is equivalent to a linear regression model with a single predictor (age category).
4. *Age and Race* - We employ the same approach as above, except we compute an average for each combination of age category and race/ethnicity category. This model is equivalent to a linear regression model where each combination of age and race/ethnicity is included as a predictor.
5. *Age, Race, and RU-13* - We employ the same approach as above, except we compute an average for each combination of age category, race/ethnicity category, and RU-13 county designation. This model represents the most complex descriptive statistics reported by Okpych (2015).
6. *Age and State*- We employ the same approach as above, except we compute an average for each combination of age category and state.

These baseline models can be interpreted in two ways. First, as noted, five of the six baseline models can be interpreted as simple linear regression, akin to those used as baseline models by Salganik et al. (2020). These regression models can then, as all regression models, be used for prediction. Second, these models can be thought of as first grouping youth along a set of demographic variables, and then taking the average number of services received for those youth as the prediction. However, these baseline models are simplified both in the independent variables used, and in the assumption of a linear relationship between these variables and the outcome.

It is also useful to compare our models to baselines that use all of the same independent variables, differing only on the linearity assumption. To this end, we include as additional baselines two regression models. The first is a linear regression model, which uses all of the same variables as our tree-based models (those identified in Table 2). Because the outcome is a count variable, we also use a negative binomial model that incorporates a set of fixed effects for each age/state combination, as one

[4] The hyperparameters we selected were: 1000 decision trees with the max depth of any tree being 15 and the minimum number of samples required to split a node being 12. We also require that the minimum number of samples in a leaf node is 5. Bootstrap samples are used to build each individual trees.

[5] The hyperparameters we selected were: 0.15 for the learning rate, zero loss reduction required for a node split, 70 percent of the features are randomly selected when building each individual trees. Minimum samples to split a node is 3, and the max depth of any tree is 6.

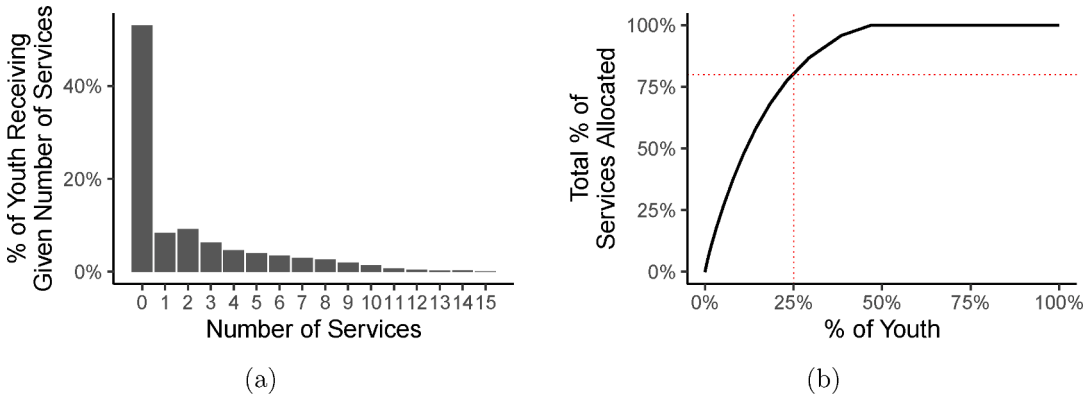


Fig. 2. A) The percentage of youth (y-axis) that received each possible number of unique services (x-axis). B) The total percentage of all unique services (y-axis) allocated to a given percentage of youth (x-axis). Red dotted lines mark the fact that around 25% of all youth receive around 80% of all unique services allocated.

would do for a traditional statistical analysis. [6]

4.2. Evaluation

Our evaluation is conducted using a procedure known as *K-fold cross-validation* (Mosteller and Tukey, 1968). We first split our data into K subsets, here, $K = 10$. We then treat the first subset as the test data and the remaining (90%) as the training data. We repeat this same process using the second subset of the data as the test data, and then the third, and so on. We therefore train and evaluate the same models ten different times, allowing us to obtain confidence estimates on our evaluation metrics.

We primarily use the *Root Mean Squared Error (RMSE)* to evaluate our model. Formally, given a set of N youth, the RMSE is computed as $\sqrt{\frac{\sum_i^N (\widehat{y_i} - y_i)^2}{N}}$, where $\widehat{y}_i$ is the predicted number of services for the ith youth in the dataset, and y_i is the true number of services that youth received. RMSE is the standard measure for evaluating predictive accuracy for models that attempt to predict a continuous outcome variable.

Motivated by the potential application of predictive modeling at the point of decision, we show in our Discussion section results from a second metric designed to evaluate potential inequalities in how services would be allocated if any of the models we used here were to be used to allocate services in a child welfare setting. To this end, we compute the *mean error*, $\sum_i^N \frac{\widehat{y_i} - y_i}{N}$, for each model, *for each race/ethnicity of youth.* The mean error provides a measure of how many services would be allocated by a given model, on average, for youth in the test data relative to the actual amount they received. A *negative* mean error would therefore mean that on average, youth of a given race/ethnicity received fewer services when allocated by a given model than they actually received. A *positive* mean error would mean that on average, these youth received *more* services if services were allocated by the model than they actually received in the real world.

4.3. Model exploration

Once establishing the most predictive model—that is, the model that "understands" the data the best—we then explore what this model can tell us about the factors that are associated with a youth receiving more (or less) services. In a traditional regression model, we can assess which variables are most predictive by simply looking at regression coefficients. In more complex models, however, or in regression models with high levels of collinearity, it becomes a challenge to determine the impact of any one factor on the predictions of a given predictive model.

To address this challenge, scholars have constructed various methodologies; a review of which can be found in Roscher et al. (2020). Here, we adopt one of the most popular methods, called *SHapley Additive exPlanation (SHAP) values* (Lundberg et al., 2017). SHAP values are quantities that can be computed for each independent variable for each youth that represent the expected change in the number of services the youth would receive, given that youth's value for the independent variable. Aggregated, or analyzed, over all youth, SHAP values can therefore give a sense of the way in which a change in a given independent variable impacts the predicted number of services a youth receives.

5. Results

We begin by presenting descriptive statistics that provide further insight into our outcome variables, and in turn, give additional context for our results. We then devote sections to describing the results of the predictive experiment we conducted, and to an exploration of the independent variables that had the most predictive power in the best performing model.

5.1. Descriptive statistics

We find that certain types of services were much more likely to be allocated than others. Fig. 1 lists each of the 15 services in the NYTD, along with the percentage of youth that received each service. Academic support and independent living needs assessments were the most frequent, with 9.6% and 7.5% of youth receiving these services, respectively. The least frequently received services were Supervised Independent Living (SIL), Room and Board Financial Assistance, and Educational Financial assistance. Only around 1.5% of youth received these services.

Similarly, we find that most youth do not receive any services, and that most services go to a small percentage of youth. Fig. 2a) shows that 54% of the youths received zero services, and 82% of youths received five services or fewer. Only 5% receive more than 8 services. Fig. 2b) shows these disparities in service allocation from a different perspective, plotting cumulative service receipt against rank. Rank represents the top percentile of service users. For example, a rank of 0.2 would represent the top 20% of youth with respect to the number of services received. We can see that the top 25% of youth received roughly 80% of the services.

These descriptive statistics align with prior work that service allocation is concentrated on a relatively small number of youth (Pérez et al., 2020). However, we also find here that they connect with more

[6] Notably, another appropriate model is a zero-inflated Poisson model. Attempts to estimate this model using our data were unsuccessful unless we limited the independent variables to a subset of those listed in Table 2; as such, we do not include those results here.

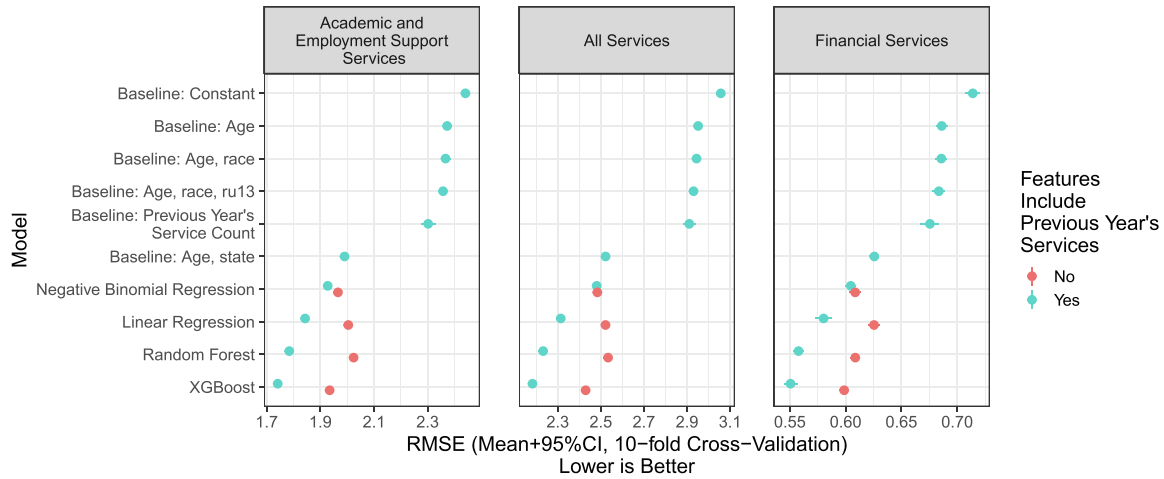


Fig. 3. Results of our predictive experiment, using Root Mean Squared Error (RMSE; y-axis) as the outcome of interest. Each row represents a different prediction model, and each of the three sub-plots shows results for the three different dependent variables we analyzed, respectively.

general findings of a Pareto Principle (the so-called "80/20 rule") in public services (Caspi et al., 2016), in that roughly 20% of youth receive the vast majority (roughly 80%) of all services.

5.2. Evaluation

We find that the more complex models we develop significantly improve over plausible baseline methods in predicting the number of services received by youth, but that this increase in predictive performance seems to draw from its ability to leverage a small number of important variables. Across all three dependent variables considered, the tree-based models using the full set of independent variables significantly outperform all other predictive models. On average, the GBT model was within 2.18 (95% bootstrapped CI of [2.17,2.20]) services of the actual total number of unique services received by youth in 2018, and within 0.55 [.54,.56] and 1.74 [1.73,1.76] for Financial Services and Academic and Employment Support Services, respectively. As a point of comparison, this is an improvement of 25.5%, 26.1%, and 19.4% in predictive accuracy on our three outcomes, respectively, as compared to the baseline model, which makes predictions based on age, race, and RU13. This baseline model, as described above, takes the average service count for each unique combination of age, race, and RU13 code in the training data, and uses those to predict service receipt in the test data. It represents the statistical models employed in the study by Okpych (2015).

Fig. 3 also shows that the predictive performance of the more complex models is statistically significantly better than baseline predictions even if we remove the most predictive feature: service count from FY 2017. However, in doing so, we see a drop in the magnitude of this improvement; from around 20% to around 5% when averaged across the three dependent variables. On the one hand, then, evaluation results give us confidence that these more complex predictive models capture important dimensions of variation in how services are allocated that may be useful in better understanding the data. Because we look at three different outcome variables, we also have confidence that this claim generalizes to various underlying reasons why services are allocated. On the other hand, these results suggest that even highly complex machine learning models may rely heavily on a small subset of all possible factors in the NYTD and AFCARS datasets. We explore the implications of this in the following sections.

5.3. Model exploration

We find that across all three outcome variables, the majority of independent variables we use from the NYTD and AFCARS have no utility in predicting the number of Chafee services a youth will receive. Results are presented in Fig. 4, which shows the distribution of average (absolute) SHAP values for each independent variable used in the GBT models across the 10 different folds of our cross-validation. More specifically, we find that 70% of the independent variables incorporated into the most predictive model impact predictions, on average, by 0.01 services or fewer. Fig. 10 in the Appendix shows that these findings are not restricted to average SHAP values, but also extend to more extreme cases. In particular, we find that even when considering (absolute) SHAP values at the 99^{th} percentile of magnitude, the estimated impact on predictions for these variables is still at or near zero. Thus on average, and even at the extremes, the model we find to have the best alignment to the data rarely makes use of the majority of variables in the NYTD and AFCARS to make predictions. In particular, none of the factors relevant to removal or to caregivers from whose home the child was removed show any measurable predictive power on the three outcome variables.

The remaining 30% of variables that do show some predictive value can be further separated into three variables that account for most of the model's predictive power, and a larger subset that provide clear but significantly smaller levels of predictive power. The larger class of variables that account for small but salient levels of predictive power consists of demographics (except for age) and a subset of case-level factors. With respect to demographics, we find that sex and urbanicity are both predictive of service receipt, although less so for Financial services given the limited number of such services allocated. In particular, female sex is associated with higher service receipt, with an average increase in prediction of .03, .06, and .11 for the financial, employment/academic, and all services outcome variables, respectively. Except for the financial services variable, for which we find no effect of urbanicity on service receipt, youth living in metropolitan areas are predicted to receive fewer services on average (approximately .1 services, on average, for the two non-Financial services categories). Regarding case-level factors, we observe that placement setting, case goal, and latest length of stay are all moderately predictive of service receipt. For the All Services condition, a SIL program placement is associated with an average increase of.39 services, and having long-term foster care as a case goal an increase of.20 services. Being labeled as a runaway or having no set value in AFCARS for placement setting or case goal are associated with average decreases in predicted service receipt of .41, .30, and .23 services, respectively. Findings for the Employment/Academic condition are similar. For the Financial services condition, only one of these factors predicts service receipt: supportive independent living placement, contributing to an average increase in services of .24. Finally, a longer length of stay predicts fewer services on average, though the effect is nonlinear - see Fig. 9 in the Appendix. In particular,

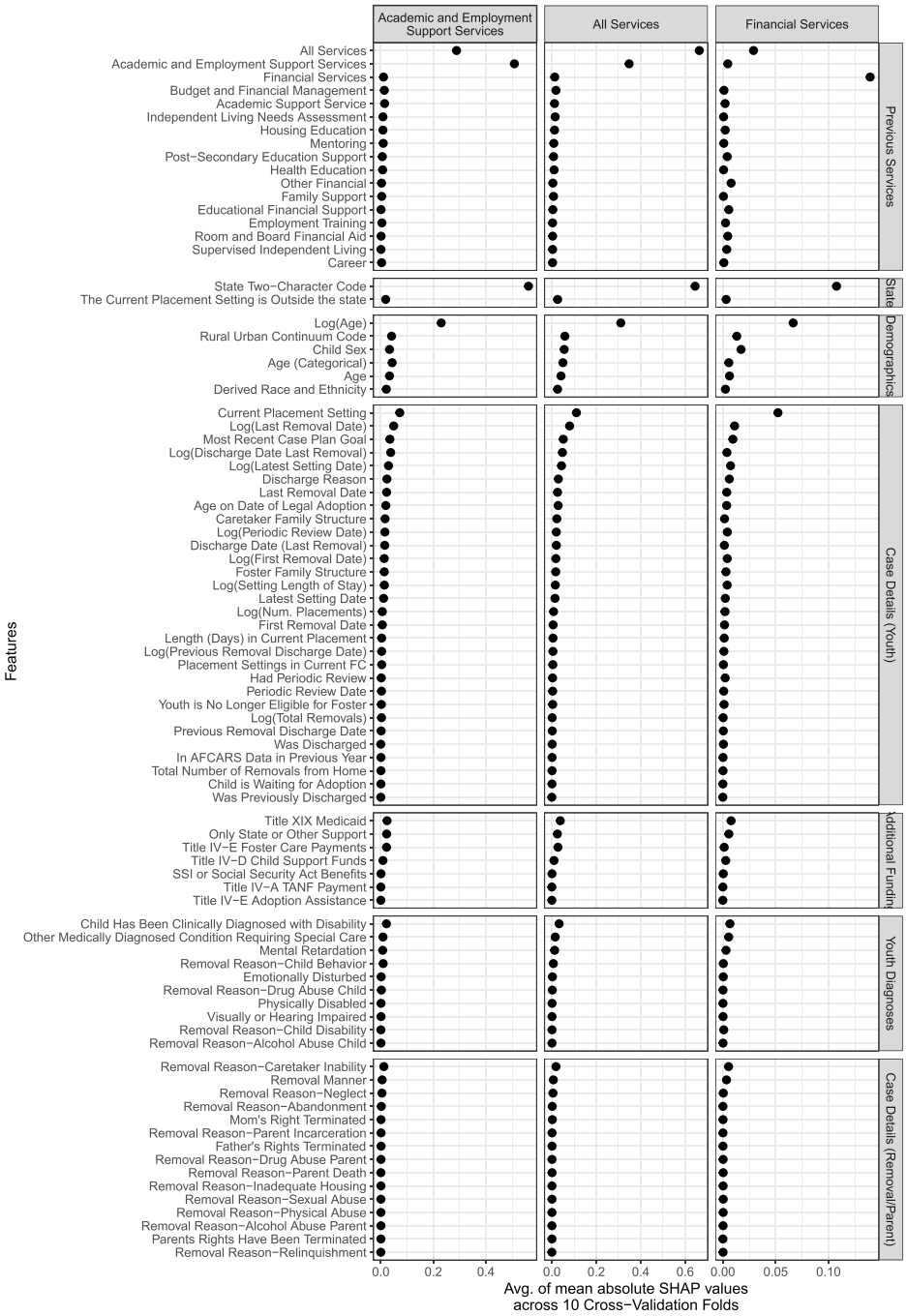


Fig. 4. Average absolute SHAP values (x-axis) for each independent variable (y-axis) for each outcome (different columns). SHAP values are aggregated across levels of categorical variables. Independent variables are also separated by high-level type (different rows of subplots). Results show mean values with 95% CIs across the ten training folds, where results for each fold are themselves averaged across each youth in the training data. Note that because youth receive significantly fewer Financial Services in general, SHAP values are smaller for this outcome variable.

conditioned on age, the odds of service receipt diminish over time, with the largest drop coming after around 4 years in care.

The remaining three independent variables that account for the majority of the GBT model's predictive power are prior service receipt, age, and state of care. Receiving services in 2017 predicted service receipt in 2018; each additional service received in 2017 was associated, on average, with a proportional increase in the number of services a youth was likely to receive in 2018. Put another way, each additional service in 2017 was associated with an additional service as predicted by the model in 2018. The association between age and service receipt exhibited similar patterns across the three outcomes, where a significant difference in predicted services receipt existed for youth who were 17 or younger, versus those older than 17 (see Fig. 8 in the Appendix). For Academic/Employment, Financial, and All Services, respectively, youth younger than 17 were predicted to receive roughly .9, .75, and .3 fewer services on average than youth older than 17.

Finally, while prior work has acknowledged the importance of state-level variation, little has been done to explore this variation. Our findings shed some light on how service allocation varies across states. Specifically, Fig. 5 provides a spatial visualization of SHAP values for each state for a GBT model for the All Service outcome; results are qualitatively similar for the other two outcomes and are thus not

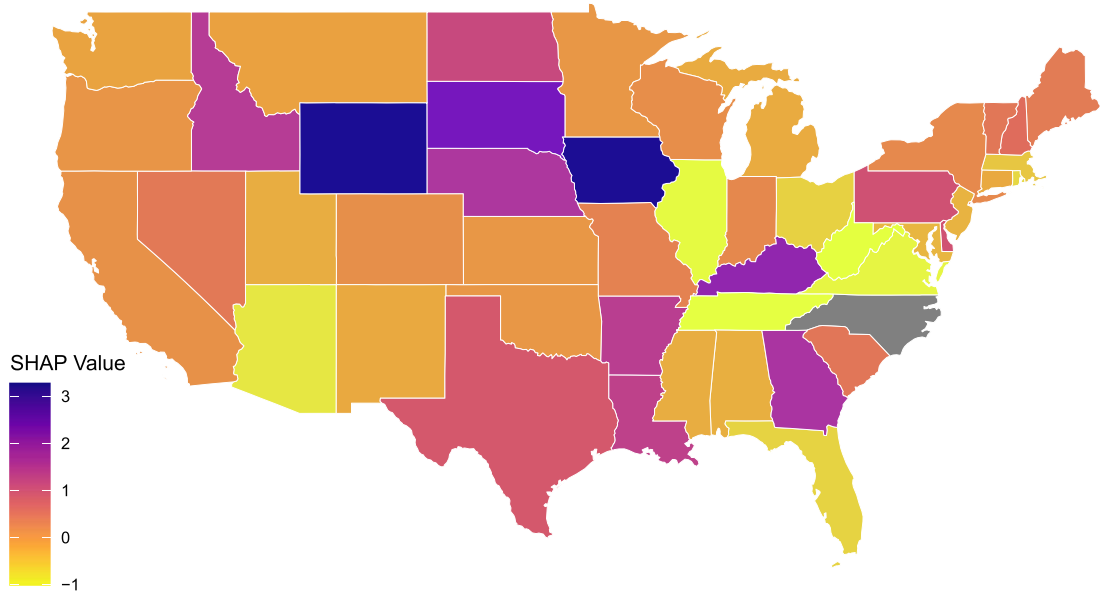


Fig. 5. The mean SHAP value (color) for youth residing in a given state.

visualized. The figure shows salient trends across some groupings of neighboring states, highlighting spatial clustering of states with high SHAP values:

1. VA, WV, and TN, where youth received fewer services on average;
2. CA and OR, neither of which had much of an effect on the number of services youth received;
3. ME, VT, and NH, all of which have SHAP values indicating a slight increase in the number of services youth received;
4. LA and AR, all of which have SHAP values that point to a moderate increase in the number of services youth received on average; and
5. WY, SD, and IA, the three states with the highest SHAP values.

A discussion of the implications of our findings for these three factors (prior service receipt, age, and state of care), and the identification of these three broad classes of variables (no predictive power, moderate predictive power, high predictive power) is provided in the following section.

6. Discussion

Our work has implications in the context of Chafee service allocation, use of the NYTD and AFCARS, and the use of predictive modeling in child welfare. In this section, we discuss these implications.

6.1. Implications for Chafee service allocation research and practice

In this section we highlight three points that are tied together theoretically by a need to understand service receipt from a systems perspective that accounts for the intertwined and iterative nature of interactions between individuals, organizations, and policies that create larger disparities through systemic feedback processes (Keddell and Hyslop, 2020). First, we emphasize that the majority of features have limited predictive power over service allocation, even when considering the possibility of higher-order interactions and including critical structural factors known to contribute to systemic inequality. Theoretically, a systems perspective explains how factors with small effects for individuals can, over time, create structural inequality (Du et al., 2022b). Our second point emphasizes this over-time nature of service allocation, describing a "rich-get-richer" effect where the best predictor of service receipt is prior service receipt. Finally, we emphasize a systems-theoretic perspective in discussing spatial clustering of service allocation.

6.1.1. Most demographic and case factors in AFCARS and NYTD have limited predictive power

Datasets as large as the NYTD and AFCARS provide the opportunity to explore a diverse array of research questions. Because of this, it can be useful to know which factors can, or cannot reliably predict outcomes. These may be factors that we can prioritize for deeper exploration, in theory or for practical purposes. The present work aids us in this sense by showing that 70% of the variables in AFCARS impact predictions by fewer than 0.01 services. This is true on average (Fig. 4) and in extreme cases (Fig. 10), and thus, most likely regardless of modeled interactions with other features. While we cannot rule out the possibility that these factors are actually predictive and our work simply failed to detect an impact, GBTs are known to be the most effective in performing a host of prediction tasks (Grinsztajn et al., 2022), and SHAP values are a state-of-the-art approach to identifying predictive features in such models. Current methodological work suggests it is unlikely that these factors would be salient predictors using other approaches.

Our work also identifies a number of factors with small but clearly non-zero predictive power: in particular, variables pertaining to sex, urbanicity, case goal, placement setting, and time in care. These *small* effects can indeed be important when we take a theoretical lens that privileges a systems perspective. Child welfare is a complex socio-technical system (Saxena et al., 2022), where small but salient effects can add up to massive disparities, e.g. across racial lines (Du et al., 2022a). Unfortunately, how best to differentiate between a variable that has a "small but critical" impact and one that has a null effect in a complex, dynamic social system is an open challenge in the field of computational social science (Hofman et al., 2021). Future work unifying both computational methods and social theory, e.g. through simulation modeling Du et al. (2022a), is critical in helping to shape our understanding of which factors emerge as important predictors when we consider effects over time.

Combining our findings with prior work can, however, inform what we should expect to find in this future work. First, prior work finds that female sex and living in a metropolitan area are associated with higher and lower service receipt, respectively (Okpych, 2015; Pérez et al., 2020; Chor et al., 2018), but extend these findings by demonstrating that age is far more predictive than either of these two factors. We discuss the

implications of age as a strong predictor in the next subsection. With regard to sex, Okpych (2015) found that higher service receipt might be related to higher college enrollment and lower incarceration rates among female as compared to male youth. When measuring recent school enrollment (as opposed to current), they confirmed these existing disparities between male and female foster youth (Okpych, 2022). Second, regarding urbanicity, it may be that youth in more densely populated areas receive less individualized attention due to how services are distributed across regions within a given state. Based on a survey of public child welfare documents and interviews with state officials, Pergamit (2013) reported that some states choose to distribute room and board funds equally across counties and cities without accounting for differences in population density. Although we did not find an effect of urbanicity on financial services allocation, it is plausible that funding related to staffing and resources are distributed in a similar manner, resulting in individual youth who live in denser metropolitan areas receiving less attention from staff and, thus, fewer services on average.

Case goal and placement settings have natural explanations; in particular, a youth placed in SIL is of course more likely to have received this service. As we explain in the subsections below, SIL programs build case management and linkage to services into the service provision framework and oftentimes require youth to engage in education, employment, or vocational programs to remain in the program (Pergamit, 2013). Similarly, youth who have run away or are otherwise missing data from their file likely represent subgroups of youth who are not able to be located or have slipped through the cracks in some way and are, therefore, either less likely to receive services or more likely to have improperly maintained files. In other words, it is possible that youth with missing data received more services than were actually documented.

No prior work has yet explored the relationship between length of stay in care and services received. Past research has posited that eligible youth who are in foster care extend their stays to leverage more services, which may result in a correlation between extended foster care and the number of services (Courtney et al., 2014). Notably, in the Midwestern Study (Courtney et al., 2011), transition age youth reported that the quantity of services is consistent with length of stay. On the one hand, we find the opposite: conditioned on age, the longer youth spend in care, the fewer services they receive on average. On the other, we note that (1) this effect is conditioned on youth age, and (2) the effects of time in care are limited overall, and even further limited beyond the binary of whether a youth was in care for more than one or two years, or not. Our work therefore poses a new avenue for research exploring the role of length of stay, contingent on age and other factors, in determining Chafee service allocation.

6.1.2. *A rich get richer effect?*

In addition to showing where not to look, or where to look with the lens of systems thinking, our work also shows where we must look if we are to understand the bulk of variability in Chafee service allocation. In particular, the vast majority of our ability to predict Chafee service receipt from the NYTD and AFCARS datasets can be attributed to only three variables: how many services the youth received previously, how old the youth is, and in which state the youth resides. Our finding regarding age is not surprising, as youth at age 17 are both preparing to transition out of care, and reaching milestones with regard to employment or post-secondary education.

Of particular interest in this section is that conditioned on being of a particular age in a particular state, by far the most important factor in determining service receipt is simply whether or not youth are already receiving services. This finding underscores the importance of understanding the factors that predict first-time service receipt, a population explored by Chor et al. (2018). Future work should explore this avenue with the methods utilized here. However, we note that Fig. 3 suggests that it may be difficult to make such predictions beyond what can be predicted using age and state. This is because the figure shows that models with all variables except prior service receipt show limited improvements over a baseline of age and state alone.

Still, there are theoretical implications of our findings even without analyzing initial service receipt. In particular, Chafee service allocation appears to present a kind of *rich get richer* effect, both within a given year (where 80% of services go to 20% of youth) and across years (where the same youth who received services last year received them again this year). One possible explanation is that youth who receive some types of services are required to engage in other services as a part of the eligibility criteria. For example, youth who receive room and board funds are frequently required to work on a budget and maintain employment or academic engagement (Pergamit et al., 2012). The former is a service in and of itself and the latter would likely be associated with receipt of academic or post-secondary educational support, career preparation, and employment programs or vocational training. Thus, the rich may get richer solely because they are required to do so to maintain their eligibility for one or two other services. A second possible explanation is that caseworkers, structural conditions, and/or policy may, implicitly or explicitly, dichotomize youth into those who should receive services, and those who should not. As noted, however, what exactly these characteristics are does not seem to be captured by the NYTD or AFCARS, or these factors would have emerged as salient predictors. In a related vein, a final explanation may be that this rich-get-richer effect is determined by how data are reported to NYTD and AFCARS, with systematic (and unobserved) differences in how services for certain youth are reported in contrast to others. Future empirical and theoretical work is critical to disentangling these lines of inquiry, but our current study showcases the utility of a forensic social science approach in identifying starting points.

6.1.3. *Spatial clustering*

Finally, there are several potential explanations and points of departure regarding our finding of spatial clustering in service allocation rates. First, theories of interjurisdictional policy diffusion and budget spillover point to many reasons that government spending in one state might influence that in another, such as concerns about appearing too generous or austere relative to other states, or constituents' tendencies to compare state politicians with those in neighboring states (Baicker, 2005). In the case of medical spending, Baicker (2005) identified interstate migration as having the most influence on budget spillover between states. They stressed the importance of evaluating a range of neighborliness metrics, including geographic proximity, similarities in demographic composition, per capita income, and population size. Exploring these variables could help identify possible causes for clustering between contiguous states.

Second, prior research has revealed interdependence between neighboring states whereby one state's economic growth is dependent on productive government spending in another; conversely, budget cuts in one state can have detrimental effects on others within proximity (Ojede et al., 2018). Youths' needs for room and board, other financial assistance, and education and training vouchers might be influenced by broader economic trends related to the housing market, labor market, and public higher education system, all aspects of a state's economy that are likely dependent on the economies of nearby states. For instance, Ojede et al. (2018)'s (2018) findings demonstrated how public spending on higher education could contribute to growth in per capita income within neighboring states. Equally important are common economic factors, such as regional economic downturns, that may impact spending across several contiguous states in a similar manner (Baicker, 2005). Policy diffusion is yet another area worthy of exploration. In the past, states have exhibited regional clustering patterns related to the length of time it took to adopt child welfare policies based on federal mandates (Lloyd Sieger and Rebbe, 2020).

Although we identified what looks like regional clustering between neighboring states, it is crucial to recognize that contiguity is not the only variable that plays a role in the policy diffusion and that states

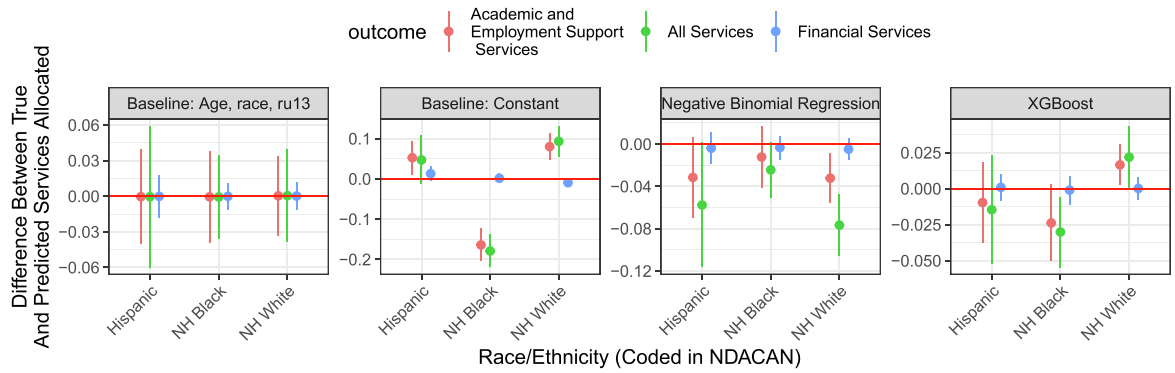


Fig. 6. Results of our predictive experiment, using Mean Error (y-axis) as the outcome of interest, with results differentiated by race/ethnicity (y-axis; only a subset of race/ethnicities coded in NDACAN are used here). Each separate subplot represents a different model (only a subset of models are plotted here) and colors represent the different independent variables considered.

outside these clusters with similar SHAP values (e.g., NV and VT) should be evaluated for potential similarities in policies and regulations (Karch, 2007/ed). For example, policymakers in one state may imitate another state's policy initiatives based on shared political commitments or similarities in demographic composition (Karch, 2007/ed), or philanthropic foundations may endorse policy diffusion across states and play a role in determining how policies are implemented once passed (Bushouse and Mosley, 2018). In these cases, the spread of policies may have more to do with the networks established by politicians and interest groups than with geographic proximity (Bushouse and Mosley, 2018). Finally, it is worth examining the influence of internal or external events that may have differentially shaped state policy environments over the past couple of decades in ways that caused long-lasting effects on service allocation in the following years (Jennings et al., 2020).

6.2. *Implications for predictive modeling*

The predictive models we develop in this work significantly outperform a number of strong baseline models on a standard evaluation metric (RMSE). As with methods of evaluating fit for more traditional statistical approaches (e.g. chi-squared tests of successively complex regression models), this evaluation shows that certain models we construct better capture variation in the data than others, and thus provide a better explanation for the underlying processes. It is clear, however, that these complex models are not always right in that they cannot perfectly predict the services youth receive. In both traditional statistics and forensic social science, we typically do not concern ourselves with the inability to *fully* predict or explain an outcome. Instead, our interest is only in 1) comparing the models we develop, and 2) using the most appropriate model to understand relationships between dependent and independent variables.

At the same time, it is instructive to consider the reasons why our predictions are imperfect, and what the implications of these imperfections might be. This is perhaps especially true in the context of predictive modeling at the point of decision, where prediction error directly translates to misguided action. Computational social scientists have explored various causes for this so-called limit of predictability, (Song et al., 2010; Martin et al., 2016) (Hofman et al., 2021); we review a number of these here in the context of our work and discuss their implications for predictive modeling in child welfare more broadly.

6.2.1. *Bad Models and/or Bad (Operationalizations of) Data*

One reason we cannot perfectly predict outcomes may be that the predictive models we use are bad or wrong. To this end, much of the computational work in the field of human services is targeted at constructing better and more predictive models (Saxena et al., 2020b). While we cannot rule out the possibility that these models could guide us to better predictions, there is limited evidence that these efforts drastically increase predictive power.

What does seem to impact predictive capabilities is data quality. In most areas of machine learning, better data tend to lead to better predictions (Jarrahi et al., 2022). One critical way in which data can be poor is if variables are poorly operationalized. In Section 6.3, we note several ways in which both our own decisions on operationalization, as well as decisions made in the NYTD and AFCARS, could have lessened data quality. Others, too, have noted challenges to operationalization in the NYTD that complicate its use for important questions (Okpych, 2022). In addition to operationalization, however, are limits on what variables are selected for inclusion in the model. That is, while not always the case, machine learning models generally benefit from the inclusion of additional variables.

In our case, however, it is not immediately clear what these additional variables might be, as we use virtually all available data in AFCARS and the NYTD for our predictions. Indeed, we even leverage several independent variables from 2018 to make predictions about the same time period. While this does not hamper our ability to study factors associated with service receipt, it does imply that even when leveraging case factors in existence at the same time as our outcome variable, predictive models built using the NYTD and AFCARS cannot fully predict outcomes. Point-of-decision predictive modeling may therefore have similar limits of predictability when data consist of relatively generic, high-level information about youth like in the NYTD and AFCARS, and in administrative data more generally.

The knee-jerk response to limited data quality and/or quantity is that to create better predictive models, we must therefore collect more and/or better data. However, more fine-grained data raise significant ethical and equity questions with respect to surveillance (Abdurahman, 2021a). The limits of predictability therefore lay bare that there is *potentially* a direct value trade-off between surveillance and predictive capacity. We say "potentially" because until such data are collected and predictive models built on top of it, we cannot be certain that the data's existence will improve our predictions. Assuming more data does mean better predictions, Abdurahman et al. (2021b) reminds us that we must ask what the value of better prediction is relative to the well-established dangers of surveillance.

We note, however, that Abdurahman et al. (2021b) largely considers the dangers of surveillance in terms of youth and families, in contrast to surveillance of the system itself. For example, noticeable in NYTD and AFCARS are gaps in our ability to link youth to their service contexts beyond high-level indicators of state and, in some cases, zip code. Given the central role of states in our predictions, and if increased predictive capacity is the goal, we suggest that more significant consideration is given to making data on service providers available to the public. This would help us address questions like, what individual case

characteristics are most salient in determining the need for particular classes of caseworkers and/or agencies? To what degree do these perceptions of need match real needs? How might personal or organizational biases shape these perceptions and the decisions made based on them? What criteria are considered when determining which youth receive which services, what constraints do caseworkers face in exercising their discretion, and what features of the service delivery system are unique to the state, county, or agency where the decision is being made? These are lines of inquiry and data that can help us elucidate the quality, quantity, and nature of available services and estimate with greater precision how many young people receive individual services within each category.

6.2.2. Good predictions for some, bad predictions for others

Limits of prediction may also emerge if we are able to make effective predictions for some subgroups of youth, but not others. Fig. 6 presents results for our mean error metric and shows, for example, that the most accurate model in our work does indeed make systematically biased errors when sub-dividing youth by Race/Ethnicity. This finding surfaces a distinction in the ways we think about the limit of predictability and its associated causes in the context of forensic social science versus point-of-decision predictive modeling.

In the context of forensic social science, our models are slightly miscalibrated, meaning that our estimates of model accuracy differ slightly based on race/ethnicity, as do our estimates of feature importance using SHAP values. While these differences are worth noting, they are small enough that they would not substantially change the interpretations presented above. On the other hand, if the GBT model presented here were to be used at the point of decision, it would provide more services to non–Hispanic White youth than it would to non–Hispanic Black youth, or to Hispanic youth for two of the three outcomes considered. This is evidenced by the fact that the mean error for Hispanic and Non–Hispanic Black youth are significantly lower than for Non–Hispanic White youth for the GBT model. In contrast, the simple descriptive model developed by Okpych (2015), as well as the non-tree-based negative binomial model we assessed here, would show no such favoritism.

Finally, our comparison of predictive models serves as an exemplar of evaluations that prioritize equity as a value. Equity is one of three criteria, along with ethics and bias, known to be relevant to child welfare but lacking sufficient attention in peer-reviewed studies using machine learning to predict child welfare outcomes (Hall et al., 2023). Even though we do not suggest any concrete applications of our findings to practice settings, we consider how using our best performing model to guide decision-making might generate an "inequitable distribution of resources based on sociodemographic characteristics" (Hall et al., 2023, p. 7). Future research should continue to explore the roles of equity, ethics, and bias in contributing to patterns of service allocation within the child welfare system. We stress the importance of establishing methods for testing the possible outcomes of algorithmic decision-making in child welfare based on identified criteria before even considering how these models might shape decisions in a particular area, such as service allocation for older youth. Understanding how predictions can go awry and showing how to detect signs of these problems in models moves the field forward in achieving this goal.

It is the case here, as has been shown elsewhere in child welfare (Chouldechova et al., 2018), that more predictive does not necessarily mean more equitable. It is worth noting that a significant amount of work has sought to address these inequalities in the predictions of machine learning algorithms (Mitchell et al., 2021). And if equity is considered a priority throughout model development, it is possible to ensure high levels of accuracy without the kinds of inequalities appearing in the models developed here (Rodolfa et al., 2021).

6.2.3. The role of randomness and child welfare expertise

Finally, using the best possible model with all possible data, one could imagine prediction errors may emerge simply from the fact that human behavior always introduces some level of randomness (Martin et al., 2016). Of course, because we will never have all the data we might want, nor can we ever know whether there exists a better modeling approach than the ones we can imagine, we can never truly know the role that randomness plays relative to what we can control (Hofman et al., 2021). As such, there is no clear answer to the question, "how good of a predictive model is good enough to use?" This holds true for applications in both forensic social science and at the point of decision. In each case, then, it is critical that we safeguard model predictions with domain expertise. In particular, in both cases, existing theory and practice and careful consideration of predictions by domain experts are both critical in ensuring appropriate use of predictive modeling (Saxena et al., 2020b).

6.3. Limitations and avenues for future work

Our work contains a number of limitations that should be considered when assessing our findings. With respect to the challenges to operationalization noted above, it is questionable whether a service delivered in one setting can even be compared to the same service delivered in another setting. For instance, the NYTD definition of career preparation is broad enough to encompass anything from a five-minute talk about the importance of professionalism to an intensive program that provides guidance and support at every step of the job search and retention process. Not only are there bound to be variations across jurisdictions in local norms and procedures governing how these services are defined, delivered, and documented as data, but across and within agencies, as well. Whereas one agency might have a detailed guide outlining the requisite components of each service such that the activities involved are relatively standardized, others might employ a vague definition that makes no distinction between a five-minute conversation and a comprehensive program, meaning that two youth at the same agency might receive entirely different services documented under the same heading. A 2017 Child Trends survey of Independent Living Coordinators across states found that some states checked several boxes under each service umbrella, while others checked very few for any of the domains; the use of evidence-informed and evidence-based interventions was similarly variable across states (Fryar et al., 2017).

Just as important is that service receipt is not a direct indicator of need and is always at least partially reliant on youths' existing connections to agencies that either deliver these services or refer youth to them. Our finding that youth in SIL programs received more services on average is illustrative of this phenomenon. Such programs are often required by the state to provide the full menu of independent living services to all participants by default (e.g., NY Requirements for Each Supervised Independent Living Unit, 2015). Conversely, we found that youth who had run away from their placements received fewer services, most likely because the system failed to maintain consistent contact with them throughout the duration of the reporting period. The recipients for any given service likely only represent a fraction of the total eligible population in need of that service and may just be a subset of program-engaged youth identified by individual staff as deserving of service receipt, whether they needed it more than other participants or not. Clearly, there is ongoing need for interventions designed to rectify disparities in key outcomes between foster youth and the general youth population. Less than five percent of youth in our study received education financial assistance and post-secondary educational support, yet there remain substantial disparities between foster care alumni and youth without foster care histories in rates of post-secondary education completion (Gypen et al., 2017). This means that the academic and educational financial support currently being provided to youth has not been effective in offsetting the barriers they face to educational attainment.

More specific to our study, our analysis predicted counts of services categories, which say nothing about the quality of each service. Doing so

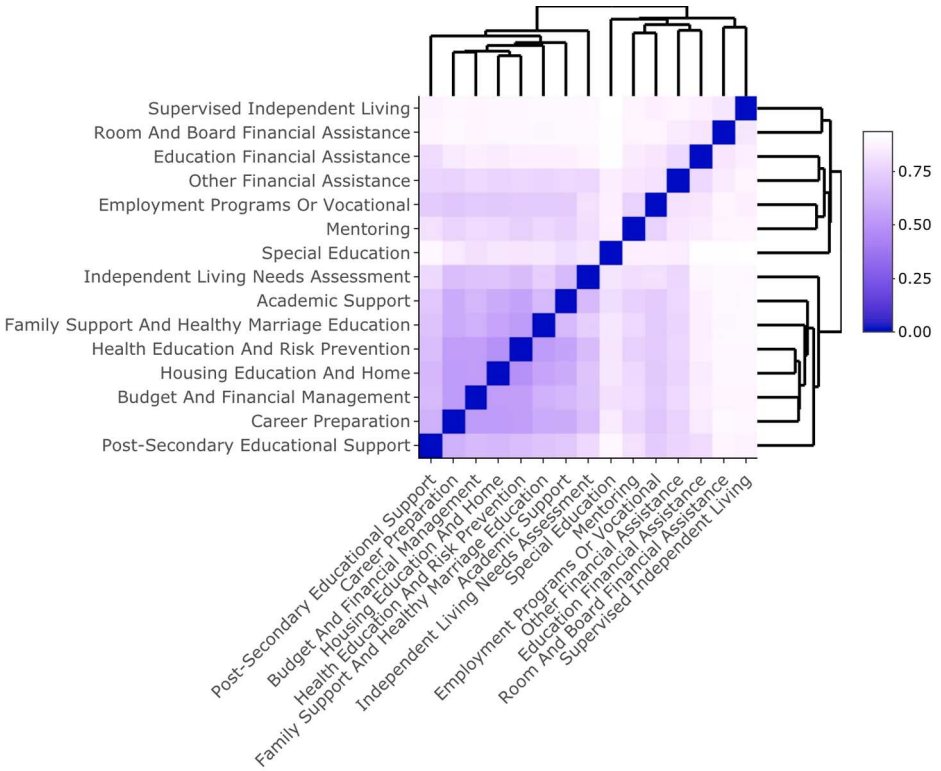


Fig. 7. Distance, based on 1 minus the Jaccard coefficient, between each pair of Chafee services. Each cell represents the value of one minus the Jaccard coefficient (described in the article body) measured between a pair of Chafee services. The higher Jaccard coefficient (and thus, the lower the distance based on Jaccard), the more blue the cell, and the higher the overlap in the youth who received both of those services. Services are also clustered using a hierarchical clustering algorithm, dendrograms are displayed that show these clustering patterns.

also restricts us to the assumption that the importance of obtaining one additional service is consistent no matter how many services a youth has received. Thus, for example, our model assumes the importance of going from no services to one service is the same as going from ten services to eleven. Given the inequities in how services are allocated, it seems possible that this assumption is invalid, and that more care may be needed in differentiating between the various categories of service receipt studied in future work (Chor et al., 2018).

7. Conclusion

Presently, there is robust debate in the child welfare literature about whether computational methods can help address the risks and needs of youth involved in the child welfare system. One major concern is the fear that algorithmic decision-making may appear more neutral, but amplify bias embedded in the data.

Indeed, we demonstrate that algorithms that maximize for fit may decrease equity in service allocation problems. As we argue, however, machine learning, and predictive analytics more broadly, need not only be applied in a way that makes (biased) decisions. They can also be used, for example, to (re-) illuminate these inequalities (Abebe et al., 2020), and as we show here, to help us better understand the system of child welfare, the data we use to formulate those understandings, and the criteria we use to evaluate models. By doing this, we can shed light on how different models based on various large datasets can be tested for the effects they might have on the outcome being predicted. Critically, however, absent underlying theory, the use of computational methods is likely to lead to naive interpretations of results that do nothing to advance knowledge in the field. Our work shows the benefits of a forensic social science method that interleaves predictive modeling with domain knowledge on relevant criteria (i.e. equity) and theory (i.e. systems perspective) in a mutually beneficial way.

We hope that our work encourages scholars and practitioners moving forward to understand predictive modeling and other computational tools not necessarily as tools for decision-making, but as tools for decision-understanding, and/or -evaluation (Du et al., 2022a). In doing so, we can maintain an eye not only on individualizing care, as most prior work on predictive modeling has done, but rather to emphasize the importance of macro-level factors and consider how we may intervene at this level as well. For instance, state as a predictor of service allocation raises questions about the impact of statewide policy and practice, and points to questions about funding levels, political values, and geographic service distribution, contrary to the common assumption that youth needs and demographics drive service delivery. Similarly, the associations between youth age and service delivery raise questions about whether there exists an ideal age-based timing for services and how that timing influences outcomes. In short, forensic social science strategies may help us determine which states or policies are most in need of service, versus which youth are most in need of services, which provides a new way of thinking about predictive analytics for child welfare.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are public upon request to NDACAN, all derived data (e.g. predictions) in our work are publicly available.

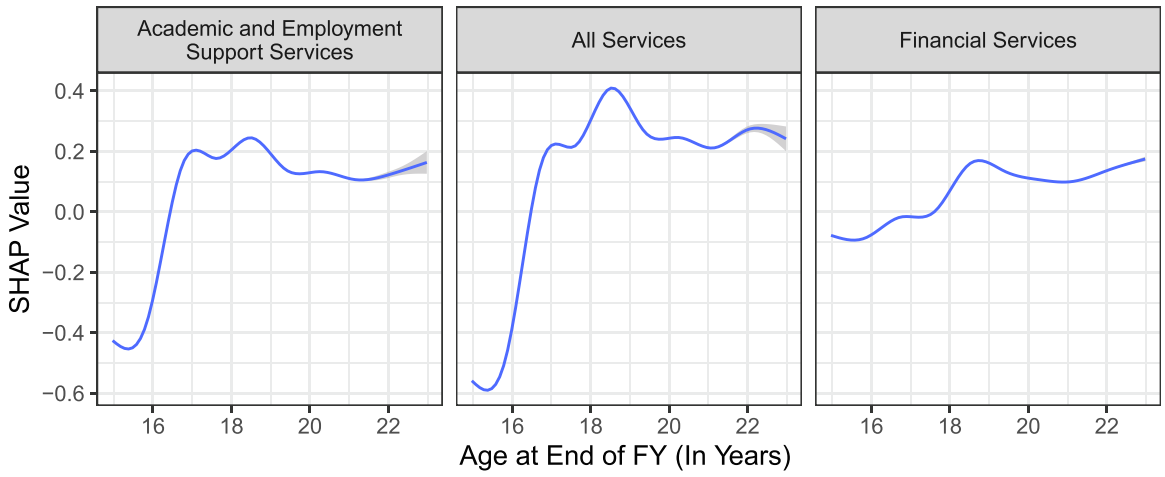


Fig. 8. SHAP values (y-axis) for the logged, continuous measure of youth age (x-axis) for each outcome variable (different subplots). Note that for the purpose of presentation, we exponentiate the value and thus the x-axis is a linear representation of age. Results are presented as the output of a generalized additive model, estimated across SHAP values output for each training point across the ten different training folds.

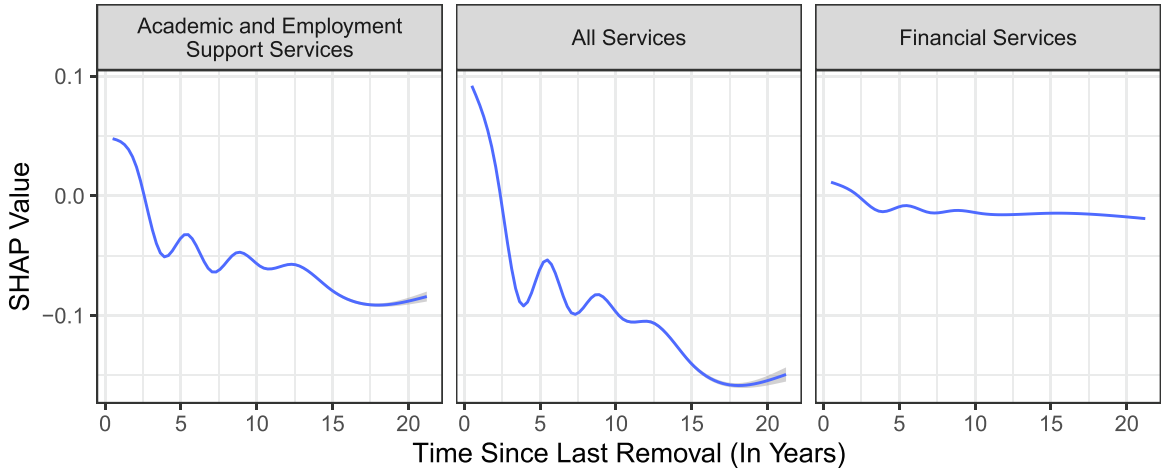


Fig. 9. SHAP values (y-axis) for the logged, continuous measure of latest date of removal from home (x-axis) for each outcome variable (different subplots). Note that for the purpose of presentation, we exponentiate the value and thus the x-axis is a linear representation of latest removal. Results are presented as the output of a generalized additive model, estimated across SHAP values output for each training point across the ten different training folds.

Appendix A. Appendix

A.1. Correlations across service categories in service receipt

In this appendix, we explore patterns in the way services were allocated across youth, providing validation for 1) the ways we split service categories into distinct outcomes, and 2) the exclusion of the Special Education service from our analysis.

Fig. 7 presents a similarity matrix in terms of distances between each pair of services based on the *Jaccard coefficient* (Niwattanakul et al., 2013). The Jaccard coefficient is a ratio, for any pair of services, of the number of youth who received both services at least once, versus the number of youth who received at least one of those services. Mathematically, this is defined as the intersection of the set of youth who received each service over their union, i.e. $\frac{S_1 \cap S_2}{S_1 \cup S_2}$, where S_1 and S_2 are sets of youth receiving two different services. Fig. 7 represents a distance matrix, where each cell represents one minus the Jaccard coefficient. If two services are given to the exact same set of youth, the Jaccard coefficient will be 1 (and thus have a distance of 0 in Fig. 7). In general, the higher the Jaccard coefficient, the stronger the overlap between the two services.

Fig. 7 provides empirical validation of decisions made when operationalizing our dependent variables. First, it shows that Special Education is an outlier, in that it has limited overlap with any other services in the set of youth to whom it is allocated. Second, Fig. 7 shows that Financial, Academic, and Employment Support Services tend to be allocated to similar sets of youth, suggesting similar underlying factors within each service type.

In addition to validating our methodological decisions, the results in Fig. 7 align with and extend prior work. Specifically, they present a slightly simplified view of the results from Chor et al. (2018) and Pérez et al. (2020). The clustering they observed across which services tend to be allocated together seems to be apparent using more straightforward methods that just address correlations between pairs of services, rather than more complex latent class analysis methods.

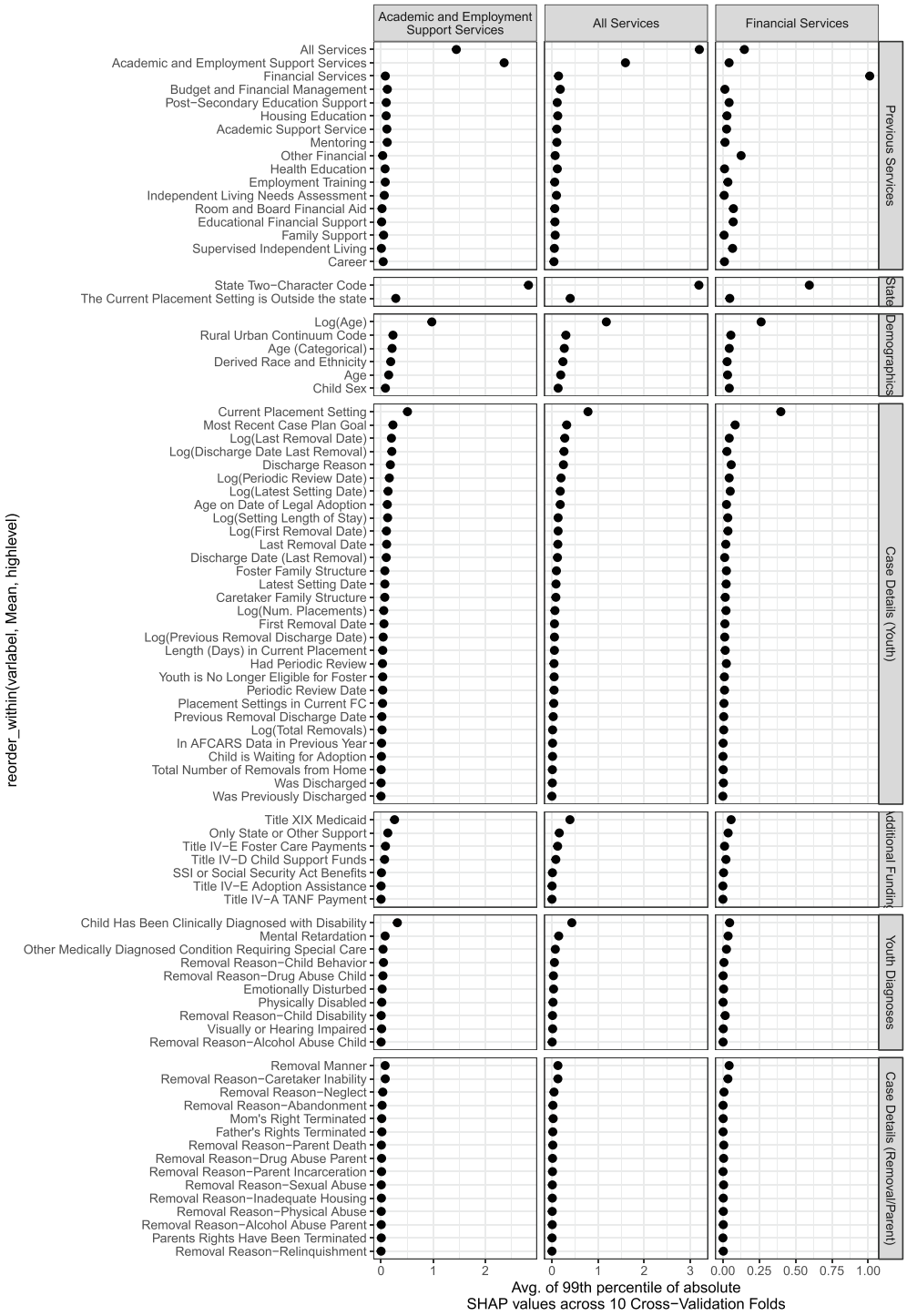


Fig. 10. Average of the 99th percentile of absolute SHAP values (x-axis) for each independent variable (y-axis) for each outcome variable (different columns). SHAP values are aggregated across levels of categorical variables. Independent variables are also separated by high-level type (different rows of subplots). Results show mean values (of the 99th percentile statistic) with 95% confidence intervals across the ten different training folds, where results are themselves computed from the distribution of SHAP values for each youth in the training data.

A.2. Additional model explanation plots

Fig. 8 and Fig. 9 present SHAP values for youth age and latest removal date, respectively, as described in the captions and the main text of the article. Fig. 10 presents the SHAP value at the 99th percentile across all youth (with mean and confidence intervals presented across the ten training folds).

References

Abdurahman, J. K. (2021a). Calculating the souls of black folk: Predictive analytics in the New York City Administration for Children's Services. *Columbia Journal of Race and Law, 11*(4), 75–110.

Abdurahman, J.K. (2021b). Calculating the souls of black folk: Predictive analytics in the new york city administration for children's services. In Colum. J. Race & L. Forum, volume 11, page 75. HeinOnline.

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D.G. (2020). Roles for computing in social change. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 252–260.

Ahn, E., Gil, Y., & Putnam-Hornstein, E. (2021). Predicting youth at high risk of aging out of foster care using machine learning methods. *Child Abuse & Neglect, 117*, 105059.

Andreswari, R., Darmawan, I., & Puspitasari, W. (2018). A Preliminary Study on Detection System for Assessing Children and Foster Parents Suitability. In *In 2018 6th International Conference on Information and Communication Technology (ICoICT)* (pp. 376–379). IEEE.

Baicker, K. (2005). The spillover effects of state spending. *Journal of Public Economics, 89*(2), 529–544.

Barker, R.L. et al. (2003). The social work dictionary.

Breiman, L. (2001). *Random forests. Machine learning, 45*(1), 5–32.

Brindley, M., Heyes, J. P., & Booker, D. (2018). Can machine learning create an advocate for foster youth? *Journal of Technology in Human Services, 36*(1), 31–36.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., and Vaithianathan, R. (2019). Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, pages 41:1–41:12, New York, NY, USA. ACM.

Bushouse, B. K., & Mosley, J. E. (2018). The intermediary roles of foundations in the policy process: Building coalitions of interest. *Interest Groups & Advocacy, 7*(3), 289–311.

Capatosto, K. (2017). *Foretelling the future: A critical perspective on the use of predictive analytics in child welfare*. Columbus, OH: Ohio State University. Kirwan Institute research report.

Caspi, A., Houts, R. M., Belsky, D. W., Harrington, H., Hogan, S., Ramrakha, S., Poulton, R., & Moffitt, T. E. (2016). Childhood forecasting of a small segment of the population with large economic burden. *Nature human behaviour, 1*(1), 0005.

Cheng, H.-F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., Perer, A., Holstein, K., Wu, Z.S., and Zhu, H. (2022). How child welfare workers reduce racial disparities in algorithmic decisions. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, New York, NY, USA. Association for Computing Machinery.

Children's Bureau, Administration On Children, Y., Families, A.F.C., Families, U.S.D.O. H., and Services, H. (2018). Adoption and Foster Care Analysis and Reporting System (AFCARS), Adoption File 2017.

Children's Bureau, Administration On Children, Y., Families, A.F.C., Families, U.S.D.O. H., and Services, H. (2019). National Youth in Transition Database (NYTD)-Services File, FY2011-2018.

Chor, K. H. B., Petras, H., & Pérez, A. G. (2018). Youth Subgroups who Receive John F. Chafee Foster Care Independence Program Services. *Journal of Child and Family Studies, 27*(5), 1402–1414.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *In Conference on Fairness, Accountability and Transparency, page*, 134–148.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134–148). PMLR.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society, 4*(2), 2053951717718855.

Church, C. E., & Fairchild, A. J. (2017). In search of a silver bullet: Child welfare's embrace of predictive analytics. *Juvenile and Family Court Journal, 68*(1), 67–81.

Collins, M. E., & Ward, R. L. (2011). Services and outcomes for transition-age foster care youth: Youths' perspectives. *Vulnerable Children and Youth Studies, 6*(2), 157–165.

Courtney, M.E., Charles, P., Okpych, N.J., Napolitano, L., Halsted, K., and Courtney, M. (2014). Findings from the california youth transitions to adulthood study (calyouth): Conditions of foster youth at age 17.

Courtney, M. E., Lee, J., & Perez, A. (2011). Receipt of help acquiring life skills and predictors of help receipt among current and former foster youth. *Children and Youth Services Review, 33*(12), 2442–2451.

Du, Y., Ionescu, S., Sage, M., and Joseph, K. (2022a). A data-driven simulation of the new york state foster care system. In 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 1028–1038.

Du, Y., Nordell, J., & Joseph, K. (2022b). Insidious nonetheless: How small effects and hierarchical norms create and maintain gender disparities in organizations. *Socius: sociological research for a dynamic world, 8*, 23780231221117888.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity, 4*(5), 41–60.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

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

Fryar, G., Jordan, E., & DeVooght, K. (2017). *Supporting young people transitioning from foster care: Findings from a national survey*. Child Trends.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644.

Geiger, J. M., & Okpych, N. J. (2022). Connected After Care: Youth Characteristics, Policy, and Programs Associated With Postsecondary Education and Employment for Youth With Foster Care Histories. *Child Maltreatment, 27*(4), 658–670.

Glaberson, S. K. (2019). Coding Over the Cracks: Predictive Analytics and Child Protection. *Fordham Urb. LJ, 46*, 307.

Goldberg, A. (2015). In defense of forensic social science. *Big Data & Society, 2*(2), 2053951715601145.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? arXiv preprint arXiv:2207.08815.

Gypen, L., Vanderfaeillie, J., De Maeyer, S., Belenger, L., & Van Holen, F. (2017). Outcomes of children who grew up in foster care: Systematic-review. *Children and Youth Services Review, 76*, 74–83.

Hall, S., Sage, M., Scott, C., & Joseph, K. (2023). A Systematic Review of Sophisticated Predictive and Prescriptive Analytics in Child Welfare: Accuracy, Equity, and Bias. *Child And Adolescent Social Work Journal*.

Harrison, T., Canestraro, D., Pardo, T., Avila-Marilla, M., Soto, N., Sutherland, M., Burke, B. & Gasco, M.A (2018). A tale of two information systems: transitioning to a data-centric information system for child welfare. In Proceedings Of The 19th Annual International Conference On Digital Government Research: Governance In The Data Age. pp. 1–2.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature, 595*(7866), 181–188.

Huang, H., Li, Y., & Campbell, J. M. (2021). Do Independent Living Services Protect Youth Aging Out Foster Care From Adverse Outcomes?. *An Evaluation Using National Data. Child Maltreatment*, 1077559521992119.

Jarrahi, M.H., Memariani, A., and Guha, S. (2022). The principles of data-centric ai (dcai). arXiv preprint arXiv:2211.14611.

Jennings, W., Farrall, S., Gray, E., & Hay, C. (2020). Moral Panics and Punctuated Equilibrium in Public Policy: An Analysis of the Criminal Justice Policy Agenda in Britain. *Policy Studies Journal, 48*(1), 207–234.

Karch, A. (2007/ed). Emerging Issues and Future Directions in State Policy Diffusion Research. *State Politics & Policy Quarterly, 7*(1), 54–80.

Kawakami, A., Sivaraman, V., Cheng, H.-F., Stapleton, L., Cheng, Y., Qing, D., Perer, A., Wu, Z.S., Zhu, H., and Holstein, K. (2022a). Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Kawakami, A., Sivaraman, V., Stapleton, L., Cheng, H.-F., Perer, A., Wu, Z. S., Zhu, H., & Holstein, K. (2022b). Why Do I Care What's Similar? Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference, DIS '22* (pp. 454–470). New York, NY, USA: Association for Computing Machinery.

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences, 8*(10), 281.

Keddell, E., & Hyslop, I. (2020). Networked decisions: Decision-making thresholds in child protection. *The British Journal of Social Work, 50*(7), 1961–1980.

Kelly, J. (2017). Illinois Drops Rapid Safety Feedback, A Predictive Analytics Tool.

Kim, Y., Ju, E., Rosenberg, R., & Farmer, E. B. M. Z. (2019). Estimating the effects of independent living services on educational attainment and employment of foster care youth. *Children and Youth Services Review, 96*, 294–301.

Kozlowski, A.C., Taddy, M., and Evans, J.A. (2018). The Geometry of Culture: Analyzing Meaning through Word Embeddings. arXiv:1803.09288 [cs].

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science (New York, NY), 323*(5915), 721.

Lee, J. S., & Ballew, K. M. (2018). Independent living services, adjudication status, and the social exclusion of foster youth aging out of care in the United States. *Journal of Youth Studies, 21*(7), 940–957.

Lemon, K., Hines, A. M., & Merdinger, J. (2005). From foster care to young adulthood: The role of independent living programs in supporting successful transitions. *Children and youth services review, 27*(3), 251–270.

Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science, 1*(1), 49–58.

Lloyd Sieger, M. H., & Rebbe, R. (2020). Variation in States' Implementation of CAPTA's Substance-Exposed Infants Mandates: A Policy Diffusion Analysis. *Child Maltreatment, 25*(4), 457–467.

Lockwood, K. K., Friedman, S., & Christian, C. W. (2015). Permanency and the foster care system. *Current Problems in Pediatric and Adolescent Health Care, 45*(10), 306–315.

Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems, pages 4768–4777.

Martin, T., Hofman, J.M., Sharma, A., Anderson, A., and Watts, D.J. (2016). Exploring limits to prediction in complex social systems. In Proceedings of the 25th international conference on world wide web, pages 683–694.

McFarland, D. A., Lewis, K., & Goldberg, A. (2015). Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *The American Sociologist*, 1–24.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163.

Mitchell, T.M. et al. (1997). Machine learning. 1997. Burr Ridge, IL: McGraw Hill, 45 (37), 870–877.
Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology, 2*, 80–203.
Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In Proceedings of the international multiconference of engineers and computer scientists, volume 1, pages 380–384.
Ojede, A., Atems, B., & Yamarik, S. (2018). The Direct and Indirect (Spillover) Effects of Productive Government Spending on State Economic Growth. *Growth and Change, 49* (1), 122–141.
Okpych, N. J. (2015). Receipt of independent living services among older youth in foster care: An analysis of national data from the U.S. *Children and Youth Services Review, 51*, 74–86.
Okpych, N. J. (2021). *Climbing a broken ladder: Contributors of college success for youth in foster care*. Rutgers University Press.
Okpych, N. J. (2022). Estimating a national college enrollment rate for youth with foster care histories using the national youth in transition database (nytd): limitations of nytd and a call to revise and relaunch. *Journal of Public Child Welfare*, 1–26.
Pérez, A. G., Harris, R. J., & Chor, K. H. B. (2020). Factors Predicting Patterns of Service Use among John F. Chafee Independent Living Services Recipients. *Child Welfare, 97* (6), 21–51.
Pergamit, M. et al. (2013). Housing assistance for youth who have aged out of foster care: The role of the chafee foster care independence program.
Prince, D. M., Vidal, S., Okpych, N., & Connell, C. M. (2019). Effects of individual risk and state housing factors on adverse outcomes in a national sample of youth transitioning out of foster care. *Journal of Adolescence, 74*, 33–44.
Purdy, J. and Glass, B. (2020). The pursuit of algorithmic fairness: On correcting algorithmic unfairness in a child welfare reunification success classifier. arXiv preprint arXiv:2010.12089.
Radford, J., & Joseph, K. (2020). Theory In, Theory Out: The uses of social theory in machine learning for social science. *Frontiers in Big Data, 3*, 18.
Roberts, A. R., & Brownell, P. (1999). A Century of Forensic Social Work: Bridging the Past to the Present. *Social Work, 44*(4), 359–369.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.
Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence, 3*(10), 896–904.
Rodriguez, M. Y., & Storer, H. (2020). A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data. *Journal of Technology in Human Services, 38*(1), 54–86.
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *Ieee Access, 8*, 42200–42216.
Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C.E., Al-Ghoneim, K., Almaatouq, A., Altschul, D.M., Brand, J.E., Carnegie, N.B., Compton, R.J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B.J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A.C., Pentland, A., Polimis, K., Raes, L., Rigobon, D.E., Roberts, C.V., Stanescu, D.M., Suhara, Y., Usmani, A., Wang, E.H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R.B., Argyle, L.P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S.P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I.M., Hummel, L.M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E.H., Leizman, B., Liu, N., Möser, M., Mack, A.E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K.M., Porter, E., Porter, K.E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J.D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M.K., Woon, W.L., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B.E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B.M., Watts, D.J., and McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences, 117(15):8398–8403.
Samant, A., Horowitz, A., Xu, K., and Beiers, S. (2021). Family surveillance by algorithm: The rapidly spreading tools few have heard of. american civil liberties union (aclu) (2021).
Sankhe, P., Hall, S.F., Sage, M., Rodriguez, M.Y., Chandola, V., and Joseph, K. (2022). Mutual information scoring: Increasing interpretability in categorical clustering tasks with applications to child welfare data. In Social, Cultural, and Behavioral Modeling: 15th International Conference, SBP-BRiMS 2022, Pittsburgh, PA, USA, September 20–23, 2022, Proceedings, pages 165–175.
Saxena, D., Badillo-Urquiola, K., Wisniewski, P.J., and Guha, S. (2020a). A human-centered review of algorithms used within the u.s. child welfare system. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.
Saxena, D., Badillo-Urquiola, K., Wisniewski, P.J., and Guha, S. (2020b). A Human-Centered Review of Algorithms used within the U.S. Child Welfare System. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, pages 1–15, New York, NY, USA. Association for Computing Machinery.
Saxena, D., Moon, S.Y., Shehata, D., and Guha, S. (2022). Unpacking invisible work practices, constraints, and latent power relationships in child welfare through casenote analysis. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–22.
Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018–1021.
Teixeira, C., & Boyas, M. (2017). *Predictive Analytics in Child Welfare: An Assessment of Current Efforts, Challenges, and Opportunities*. https://aspe.hhs.gov/system/files/pdf/257841/PACWAnAssessmentCurrentEffortsChallengesOpportunities.pdf.
Thompson, H. M., Colvin, M. L., Cooley, M. E., & Womack, B. (2021). Factors Predicting Service Referrals for Youth in the Child Welfare System. *Child and Adolescent Social Work Journal*, 1–17.
Vaithianathan, R., Rouland, B., & Putnam-Hornstein, E. (2018). Injury and mortality among children identified as at high risk of maltreatment. *Pediatrics, 141*(2), e20172882.
Yelick, A., & Thyer, B. (2020). The effects of family structure and race on decision-making in child welfare. *Journal of Public Child Welfare, 14*(3), 336–356.