Child and Adolescent Social Work Journal
https://doi.org/10.1007/s10560-023-00931-2

A Systematic Review of Sophisticated Predictive and Prescriptive Analytics in Child Welfare: Accuracy, Equity, and Bias

Seventy F. Hall[1] · **Melanie Sage**[1] · **Carol F. Scott**[2] · **Kenneth Joseph**[3]

Accepted: 7 May 2023

Abstract

Child welfare agencies increasingly use machine learning models to predict outcomes and inform decisions. These tools are intended to increase accuracy and fairness but can also amplify bias. This systematic review explores how researchers addressed ethics, equity, bias, and model performance in their design and evaluation of predictive and prescriptive algorithms in child welfare. We searched EBSCO databases, Google Scholar, and reference lists for journal articles, conference papers, dissertations, and book chapters published between January 2010 and March 2020. Sources must have reported on the use of algorithms to predict child welfare-related outcomes and either suggested prescriptive responses, or applied their models to decision-making contexts. We calculated descriptive statistics and conducted Mann-Whitney U tests, and Spearman's rank correlations to summarize and synthesize findings. Of 15 articles, fewer than half considered ethics, equity, or bias or engaged participatory design principles as part of model development/evaluation. Only one-third involved cross-disciplinary teams. Model performance was positively associated with number of algorithms tested and sample size. No other statistical tests were significant. Interest in algorithmic decision-making in child welfare is growing, yet there remains no gold standard for ameliorating bias, inequity, and other ethics concerns. Our review demonstrates that these efforts are not being reported consistently in the literature and that a uniform reporting protocol may be needed to guide research. In the meantime, computer scientists might collaborate with content experts and stakeholders to ensure they account for the practical implications of using algorithms in child welfare settings.

Keywords Systematic review · Predictive analytics · Child welfare workers · Decision making · Risk assessment · Policy

Decision-making in child welfare is a persistent topic of investigation, with debate about how to improve it going back over 70 years (Gleeson, 1987). Despite decades of work to improve child welfare decision-making, especially in areas of risk assessment and child removal decisions, research has found that decision-making remains unreliable and inconsistent. For example, Keddell (2017) found that a worker's risk aversion impacted child safety assessments and contributed to more conservative practice. A systematic review of caseworker decision-making by Lauritzen and colleagues (2018) revealed that case, worker, and organizational characteristics, as well as external factors such as policy and political climate, contributed to lack of uniformity in decision-making.

Public service systems increasingly favor data-driven performance measures and privatization of social services to cut costs and enhance efficiency (Abramovitz & Zelnick, 2015; Elgin & Carter, 2020; Huggins-Hoyt et al., 2019). Private contractors have responded to this call by offering algorithmic decision-making tools (Brauneis & Goodman, 2018; Church & Fairchild, 2017). Some have called algorithmic decision-making the next step in improving consistency and fairness in child welfare (Shlonsky & Wagner, 2005; Wilson et al., 2015). Others have warned that reliance on algorithms could render the decision-making process even more unfair (Binns, 2018), especially for families with frequent

✉ Seventy F. Hall
sfhall@buffalo.edu

[1] School of Social Work, University at Buffalo, 685 Baldy Hall, Buffalo, NY 14260, USA

[2] School of Information, University of Michigan, Ann Arbor, MI, USA

[3] Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA

Published online: 23 May 2023

exposure to public systems (Dare & Gambrill, 2017; Garcia, 2016; Gillingham, 2019a; Keddell, 2015).

To better understand these concerns, we systematically review the published academic literature on the use of machine learning algorithms for predictive and prescriptive purposes in child welfare, examining the degree to which researchers addressed ethics, equity, and bias in their methodological practices. We also explore the roles of participatory design (PD; i.e., inclusion of stakeholders in the design process) and cross-disciplinary collaboration as they relate to performance and researchers' handling of ethics, equity, and bias. We begin by providing a brief history of data-driven decision-making in child welfare leading up to the current use of machine learning, after which we present and discuss the methods and findings of our systematic literature review on algorithmic decision-making in child welfare. We conclude by elaborating upon the limitations of our review, identifying areas of future study and offering some recommendations for the future.

Data-driven Decision Making

Prior to the 1980s, child welfare agencies relied primarily on consensus-based approaches to decision-making (Shlonsky & Wagner, 2005). Consensus-based systems eventually gave way to assessments that relied on statistical risk indicators (i.e., historical data associated with higher levels of risk) to inform decision-making at different decision points. Actuarial risk assessments first appeared in the child welfare literature as early as 1984, when Johnson and L'Esperance (1984) used multiple linear discriminant analysis to predict the recurrence of physical abuse two years post-referral with 74% accuracy. Studies confirmed that these actuarial risk assessments improved inter-rater reliability (Baird et al., 1999) and predictive validity when compared to consensus-based or clinical decision-making tools (D'andrade et al., 2008). However, early actuarial decision-making researchers warned of the potential pitfalls of adopting these instruments as a mechanistic shortcut to compensate for inadequate training and resources, especially by administrators and staff who lack knowledge of statistical modeling (Wald & Woolverton, 1990). This remains an argument against predictive risk modeling today (Binns, 2018; Eubanks, 2017; Keddell, 2015).

Within the context of child welfare research, predictive risk modeling typically involves identifying a dependent variable, such as future substantiation of abuse. The focus is then to use predictive modeling (e.g., statistical methods) to identify factors that might explain or predict that outcome (Vaithianathan et al., 2013). Modern predictive risk modeling often uses machine learning to analyze thousands of cases, where each case might have hundreds of potential attributes, to identify an accurate model to predict the dependent variable. When these models are used to guide decision-making, they are referred to as prescriptive algorithms (Schwartz et al., 2017). To date, most cases of machine learning in child welfare are used for risk assessment as opposed to other potential applications, such as matching children and families to services (Saxena et al., 2020). For the purposes of this paper, the term 'algorithmic decision-making' refers exclusively to approaches that use machine learning models.

Machine learning is defined as a field of research that asks the question, "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" (Mitchell, 2006, p. 1). There is an array of subdomains within the field that study different ways that computers can "experience" the world through algorithms. The most common differentiation is between supervised and unsupervised machine learning. In supervised machine learning, researchers use existing data to examine correlations between a set of predictors (independent variables) and an outcome of interest (dependent variable) (Lanier et al., 2020). The model that encodes these correlations can then be used to make predictions about future cases based on the relationships between those variables (Lanier et al., 2020).

The three most common algorithms used in machine learning are: (1) linear models, which entail using a simple formula to find the 'best fit' *line* through a set of data points, (2) tree-based models, e.g., a decision tree or a sequence of branching operations that are yes/no, and (3) artificial neural networks, a class of algorithms originally inspired by biological neural networks (Bishop, 1995). Examples of standard methods of analysis that fall under the umbrella of supervised machine learning include standard regression models, e.g., logistic regression. Unsupervised machine learning models do not seek to make predictions, but are instead used to identify clusters and patterns that shed light on data composition (Lanier et al., 2020), for example, to group people based on characteristics, such as location, age, or gender (Dataiku, 2022). Examples of more traditional unsupervised machine learning methods include principal components analysis and latent class analysis.

The Current Systematic Review

This systematic review of the published literature aims to advance knowledge on current use cases of algorithmic decision-making in child welfare for predictive and prescriptive purposes. The current work expands on Saxena et al.'s (2020) systematic review of both statistical and machine

learning models in child welfare, which identified the need for a more robust set of predictors based on the empirical literature and advocated for the use of theory-driven modeling. Our review is narrower and more technical in its focus than that of Saxena et al. and as such, differs in a number of important ways.

First, our inclusion criteria had to be sufficiently narrow to maximize study homogeneity such that statistical tests could be conducted to synthesize findings. In particular, we restricted our review to studies that deployed sophisticated, data-driven machine learning algorithms (defined in the *Inclusion and Exclusion Criteria* subsection below) that predicted child welfare-related outcomes. By contrast, Saxena et al. used a much broader definition of algorithms inclusive of both machine learning models and older actuarial risk assessments, such as structured decision-making (SDM) tools, which are typically scored checklists consisting of current and historical factors known to predict risk (Gleeson, 1987). SDM is not a machine learning model trained and optimized on a large dataset, but a validated measure of risk much like other clinical screening and assessment tools. Moreover, Saxena et al.'s review included algorithms designed for purposes other than predicting outcomes (e.g., a goal management program and chatbot for foster youth; Brindley et al., 2018). Their thematic analysis focused on the methods used to develop algorithms, and the predictors and outcomes included within these models. They did not conduct any quantitative analyses to synthesize their findings.

Second, Saxena et al. omitted key methodological details about each machine learning algorithm, including the number of algorithms tested, performance metrics and best performing models, data sources used to train models, and the country of origin for each project, to name a few. Third, although these authors noted the crucial role of PD in integrating the needs, values, and knowledge of stakeholders and domain experts into algorithmic decision-making, they did not review the extent to which studies relied on participatory practices or cross-disciplinary collaboration. Our review aims to capture the degree to which cross-disciplinary collaborations and PD principles drive the development and implementation of algorithmic decision-making in child welfare.

Finally, Saxena et al. did not place their review within the context of recent critiques of algorithmic bias, equity, and ethics in child welfare (e.g., Dare & Gambrill, 2017; Gillingham, 2019a). These critiques tend to fall into three categories: limitations of the data sources used, problematic modeling techniques, and risks associated with how the end user uses or misuses the algorithm. Our review assesses how researchers acknowledged the limitations of the data sources they used to train their models and how they considered worker bias, racial and socioeconomic equity, and ethics in model design and evaluation. In sum, the current systematic literature review addresses the following research questions:

(1) How and where is algorithmic decision-making being used in child welfare?
(2) To what extent are these projects cross-disciplinary or participatory?
(3) To what extent do scholars address ethics, equity, and bias in their reporting of data source limitations, algorithmic design, and model implementation and performance?
(4) What factors contribute to the performance of models and the degree to which scholars address ethics, equity, and bias in their reporting of data source limitations, algorithmic design, and model implementation and performance?

Method

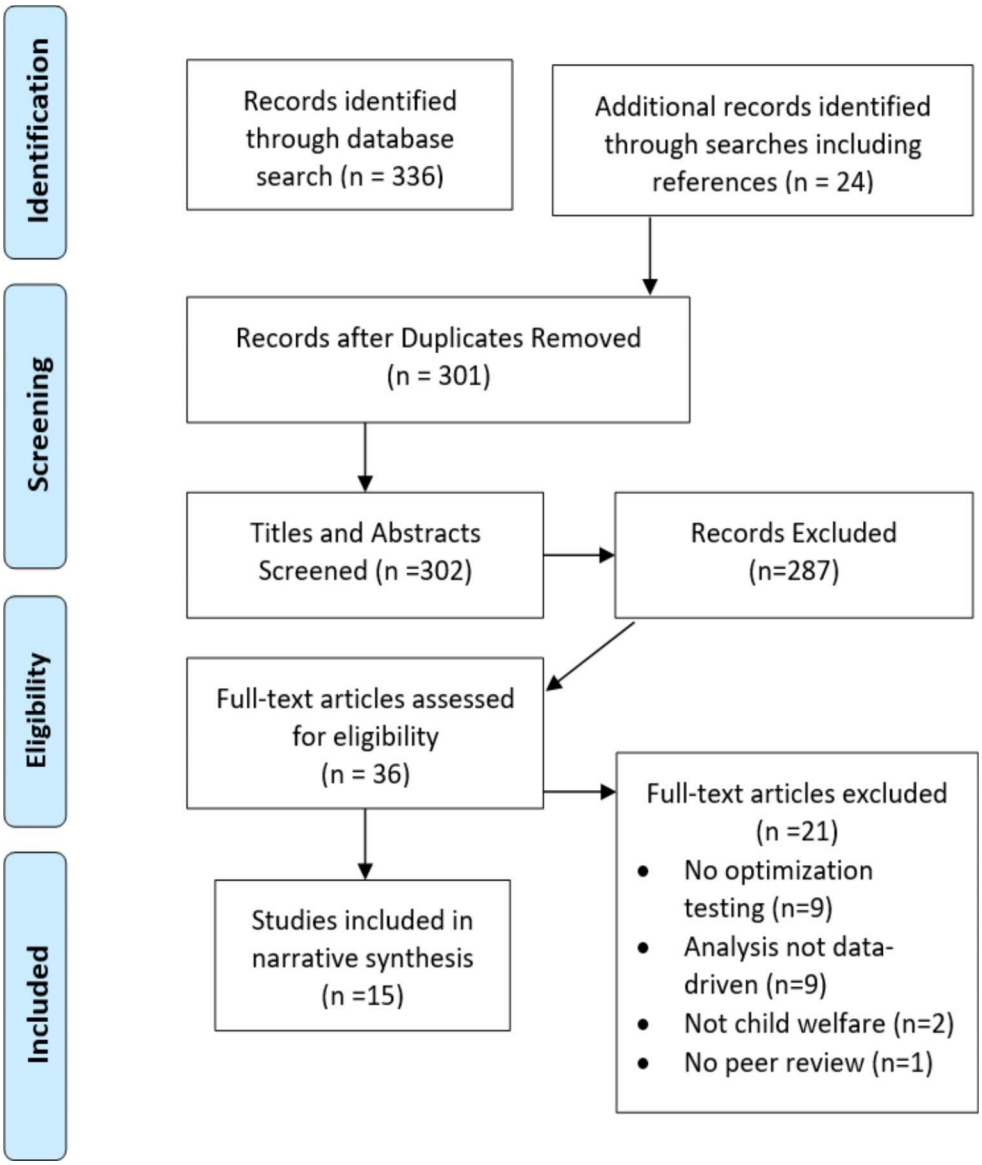
Inclusion and Exclusion Criteria

Sources eligible for inclusion within the review included English language peer-reviewed journal articles, conference papers, dissertations, and book chapters published between January 2010 and March 2020 that reported on the development and use of algorithms to predict child welfare-related outcomes and suggested prescriptive responses. We defined sophisticated predictive and prescriptive algorithms as those that used data-driven analytics and fell within the upper-right quadrant of Banerjee et al.'s (2013) *Use of Analytics in Decision-Making* chart, indicating high "analytical sophistication" and high "proactive decision-making" (p. 6). Specifically, the researchers must have at least aimed to optimize or evaluate the performance of their models using simulation or optimization techniques from the field of machine learning *and* provided recommendations for using the model in an applied decision-making setting.

Search Protocol

The first two authors designed the search strategy in consultation with a university librarian to test different combinations of keywords and identify search terms that best captured our focus on algorithmic decision-making in child welfare. We utilized a well-established review protocol, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2015) to ensure that our systematic review was rigorous and reproducible.

Fig. 1 PRISMA Flow Chart



The results of all search activities were summarized as a PRISMA flowchart, presented in Fig. 1.

We began by conducting a search of EBSCOhost Research Databases, including Social Work Abstracts, Academic Search Complete, and APA PsycInfo, using Boolean operators. An initial full-text search returned over 30,000 results, so we limited the search to abstracts only using the following Boolean query string: (*predict* and analyt* or algorithm**) *AND* (*child protection or child welfare or foster care*). We applied filters based on our inclusion and exclusion criteria. As shown in Fig. 1, this search yielded 336 results. These databases were last consulted on February 26th 2020. In addition to the above databases, we searched the Zotero bibliographies of the administrative datasets listed at the National Data Archive on Child Abuse and Neglect (NDACAN), which resulted in two additional articles. We also checked the reference lists of all screened-in sources and conducted a Google Scholar search using the keywords *algorithm OR predictive analytics AND child welfare OR foster care*. We were able to identify 21 additional sources via Google Scholar. The NDACAN bibliography and Google Scholar database were last consulted on March 15th 2020 and March 29th 2020, respectively. Finally, on August 17th 2020, one of the authors encountered an additional article while conducting a regular Google search of algorithmic decision-making efforts in different US states.

Screening Process

The first two authors scanned titles and abstracts based on predetermined inclusion criteria and narrowed the results to 36 sources, meeting to discuss and resolve any inclusion

discrepancies throughout the review process. After initial screening, one of the first two authors conducted full-text reviews to verify the eligibility of each source. As shown in Fig. 1, this step resulted in 15 screened-in texts. Among the excluded sources ($n = 21$), nine lacked optimization or performance testing, nine were not data-driven, two did not predict child welfare outcomes, and one appeared to be an unpublished paper written by a student. We reached out to the author via email to inquire about whether the paper had been published in a journal or book or presented at a conference but did not receive a reply. Because we were unable to locate these details, we labeled the paper as a document type outside of our selected criteria and excluded it from the review.

Extraction of Data Items

Data extraction was performed primarily by the first author with feedback and guidance from the second author. This process involved using a Google Spreadsheet to extract and record data elements, including descriptive information about the publication and the project, evidence of cross-disciplinary collaboration or PD, details about the researchers' methods, and performance metrics. We also noted the authors' acknowledgment of data source limitations related to ethics, equity, and bias and the degree to which the researchers integrated these factors into the development and evaluation of their algorithms. After initial data extraction, the first and second authors met to further operationalize the coding scheme for each data element. These operational definitions were largely derived from the existing literature outlining the critiques of predictive analytics in child welfare and are described in the subsections below. For additional details regarding the limitations of our screening and data extraction methods, we refer the reader to the *Limitations* section of this paper.

Project and Publication Characteristics

We recorded the year of publication, type of publication source, country where the project was conducted, and predicted outcomes for each study. Some studies sought to predict multiple outcomes, in which case we documented each predicted outcome. Examples include the likelihood of a new referral within two years of the initial referral and placement instability 18 months post-removal from the home. Finally, we recorded the purpose of the algorithms, identifying each as either applied or theoretical. Studies were labeled as applied if the model was designed to be used as a decision tool in an applied setting, such as a child welfare agency. They were labeled as theoretical if the authors made prescriptive recommendations without reporting any concrete plans for applying the model as a decision tool in a particular setting.

Methods Used

First, we recorded details about the data, including the sample size and types of data sources the researchers used to develop their models. Data sources were coded as primary data collected by the author or secondary data, such as public or private child welfare agency data, criminal justice or juvenile justice system data, medical data, public social services data (e.g., welfare benefits), private marketing research firm data, or public demographic data (e.g., census data). If the data were derived from a federally mandated source of data collection (e.g., the Adoption and Foster Care Analysis and Reporting System), the source was identified by name.

We also recorded the classification models employed by each of the studies included in the review and placed the models into one of the following categories based on a modified version of Elgin's (2018) classification scheme: (1) nonlinear classification (NLC) models, which included neural networks, support vector machines (with non-linear kernels), and generalized additive models; (2) tree-based (TB) models, which included decision trees, boosted trees, and random forests; and (3) linear models, which included logistic and linear regression models and one case of risk terrain modeling based on a binomial type II regression model. While TB is technically a subset of NLC, we found it useful to distinguish TB since it was heavily used. We also recorded counts of both the number of model categories tested (i.e., TB, NLC, and linear) and the number of specific algorithms tested across categories. For example, a study that used random forests, neural networks, and additive modeling would have tested three types of algorithms across two model categories.

Finally, we recorded information about how the model was evaluated. There are two core components to model evaluation. The first is the selection of a performance metric, that is, a numeric assessment of the quality of the model's predictions. Below, we discuss the performance metrics identified in the literature surveyed. Second, is the decision on how to divide the data into *training* and *testing* sets. This decision is crucial because machine learning models can easily be made to *overfit* - that is, to perfectly (or near perfectly) learn patterns in the training data that do not generalize to new data, or *underfit*, which occurs when a model is too simple and needs more data or training (Hawkins, 2004). Consequently, a strategy is needed to ensure that performance metrics reflect the model's ability to generalize to new datasets. Several approaches have been developed, of which two are common: (1) the *holdout method*, by which

the data are split into a train set for training and a test set on which the model is evaluated; and (2) *k-fold cross-validation*, which involves multiple splits of the data into different sets for training and testing. For more on these methods and their limitations, we refer the reader to Chap. 7 of Hastie and colleagues (2009).

Cross-Disciplinary Collaboration and Participatory Design

We explored authors' disciplinary affiliations and labeled collaborations as cross-disciplinary or monodisciplinary. We defined cross-disciplinary teams as those whose authors came from at least two of the following four groupings: computer science/machine learning (CS/ML; e.g., industrial engineering and business information systems, information technology), social data analytics (SDA; e.g., economics, epidemiology, public policy), health and human services (HHS; e.g., social work, medicine), and criminal justice. We also noted whether researchers used PD processes to develop their models. We defined studies as PD if (a) the authors reported that they consulted community members, staff, or administrators for feedback on the algorithm or (b) representatives of agencies who may use the data co-authored the article. We determined whether projects met this criterion by inspecting the authors' institutional affiliations and performing Google searches on their positions when needed.

Performance Metrics

We recorded performance metrics for all models tested in each study and assigned performance ratings to the best performing models from each study as provided by the authors. Whenever possible, ratings were based on the area under the receiver-operating characteristic curve (AUC). The AUC is a summary statistic that accounts for specificity (i.e., the portion of correctly identified negative cases) and sensitivity (i.e., the portion of correctly identified positive cases) across multiple cutoff points.

Cohen's κ and Matthew's Correlation Coefficient (MCC) were calculated for all sources that did not report AUC, but provided confusion matrices or true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values, which can be used to derive performance metrics. Both Cohen's κ and MCC measure the strength of correlations between observed and predicted classification values at specific thresholds. Whenever sources provided values at multiple cutoff points, we calculated MCC and Cohen's κ at each threshold and presented these measures as ranges in Table 1. We calculated Cohen's κ and MCC for two sources that did not report AUC (Rodriguez et al., 2019; Thurston & Miyamoto, 2018) and MCC for one source that reported Cohen's κ without AUC (Benesh, 2017). Two studies could not be rated on performance because they did not provide AUCs or sufficient information to permit calculation of Cohen's κ and MCC (Camasso & Jagannathan, 2019; Daley et al., 2016).

We rated AUC values as excellent (AUC ≥ 0.90), good (0.80 ≤ AUC < 0.90), fair (0.70 ≤ AUC < 0.80), or poor (AUC < 0.70) according to Hosmer et al.'s (2013) guidelines for evaluating model performance. As regards Cohen's κ and MCC, we relied on a simplified version of Landis and Koch's (1977) criteria for interpreting Cohen's κ, which originally consisted of six ratings ranging from "poor" (κ < 0.00) to "almost perfect" (κ = 0.81–1.00). Our ratings were as follows: excellent (MCC or κ = 0.76–1.00), good (MCC or κ = 0.51–0.75), fair (MCC or κ = 0.26–0.50), and poor (MCC or κ < 0.26). For the purposes of the synthesis (described below) performance was rated on a 4-point scale (1 = *poor*; 4 = *excellent*).

It is worth noting that MCC and Cohen's κ are not strictly comparable to AUC, nor are performance ratings comparable across models that predict different types of outcomes or that use different datasets or methods to identify training and testing sets. Although the latter limitation is unavoidable, we ameliorated the former by calculating MCC and Cohen's κ for three studies that provided both an AUC value and confusion matrices or TP, FP, FN, and TN values for specific thresholds. We also emailed authors who had not provided these values to request additional information that would permit us to calculate MCC and Cohen's κ; these attempts were unsuccessful. Ultimately, we calculated MCC and Cohen's κ for three sources that reported AUCs and did not discover any substantial discrepancies in performance ratings between these three measures. Although MCC and Cohen's κ for Wilson et al.'s (2015) study ranged from fair to good, only the first three risk percentiles were in the fair range. The remaining values were 0.53 and above. This consistency in ratings across performance metrics combined with empirical evidence that MCC and AUC values tend to be consistent (Halimu et al., 2019) bolstered our confidence in drawing comparisons between sources based on performance. These comparisons are described further in the *Synthesis of Results* subsection.

Ethics, Equity, and Bias

Next, to answer our third research question, we examined the degree to which the authors of each paper acknowledged common data source limitations across four dimensions: subjectivity, racial disparities that may be amplified in the data, the use of proxy variables (e.g., substantiated abuse as a proxy for actual abuse), and problems related to over-surveillance of marginalized communities. We scored each

source on a scale of 0 to 4 (0 = *no limitations acknowledged*; 4 = *all four limitations acknowledged*).

We also recorded the degree to which investigators took actions to ameliorate concerns related to ethics, equity, and bias. We defined ethics as a commitment to addressing the material impact of algorithmic decision-making on children and families; equity as efforts to combat surveillance, stigmatization, or inequitable distribution of resources based on sociodemographic characteristics; and bias as the influence of caseworkers' personal beliefs and attitudes on decision-making practices. Across these three categories, we coded whether the researchers (a) integrated considerations for these issues into model design or implementation (e.g., excluding variables that might reinforce racial inequality, using a strengths-based approach to model design) or (b) evaluated for the influence of these factors on the algorithm's use or performance (e.g., comparing AUC across racial groups, examining workers' use of the model for evidence of bias). This coding scheme resulted in a total of six classifications. For the purposes of the synthesis (described below) we used a scale of 0 to 6 (0 = *equity, ethics, and bias were left unaddressed*; 6 = *considerations for all three were both integrated into model design/implementation and evaluated as a performance outcome*).

Synthesis of Results

We did not conduct any meta-analyses given the small number of sources we reviewed and high degree of heterogeneity in methods and outcomes across studies (see the *Limitations* section for more details). Instead, we answered research questions 1 through 3 descriptively using counts and percentages for the following data items: (1) year of publication, type of publication source, project location, type of outcomes predicted, purpose of the algorithm, data source type, classification model type, and training/testing approach; (2) cross-disciplinary versus monodisciplinary collaboration and PD (yes versus no); and (3) acknowledgement of data source limitations and evaluation or integration of considerations for equity, ethics, and bias into model design or implementation. For question 1, we also calculated the mean, median, and range for the total number of algorithms tested across studies. We provide a complete breakdown of descriptive findings in Table 1 and summarize specific details about the data items where appropriate.

To answer question 4, we conducted several nonparametric bivariate tests. First, we calculated Mann-Whitney U tests to examine whether three variables differed significantly based on authors' dichotomous (yes versus no) disciplinary affiliations (i.e., CS/ML, SDA, HHS scholars) and whether or not the team was cross-disciplinary or incorporated PD principles into the design and implementation of their algorithms: (1) performance ratings for the best performing algorithms, (2) number of data source limitations acknowledged by the authors, and (3) the ethics, equity, and bias scale. Mann-Whitney U tests were chosen because they make no distributional assumptions (McElduff et al., 2010) and are widely considered superior to parametric alternatives when the data are ordinal and/or the sample size very small (< 10 per group), especially when the two groups are unequal (Adusah & Brooks, 2011; Simsek, 2023; Weber & Sawilowsky, 2009). The results of all three sets of Mann-Whitney U tests are presented in Table 2. Second, we calculated Spearman's rank correlations between performance rating, and the sample size, number of model categories tested, and total number of algorithms tested across model categories.

Whenever performance and sample size were included as variables, we excluded studies that lacked performance ratings and whose sample sizes were based on US states or city blocks instead of individuals ($n = 2$; Camasso & Jagannathan 2019; Daley et al., 2016). Although we chose only those variables we thought would be impactful, the novelty of the research questions and heterogeneity between populations sampled and methods used precluded us from developing a priori hypotheses for any of the statistical tests. Thus, we present two-tailed p-values for all tests conducted. We also applied a Bonferroni correction to adjust for multiple comparisons for each of the three sets of Mann-Whitney U tests and calculated confidence intervals for the Spearman's correlations using the bias-corrected and accelerated (BCA) bootstrap method (2,000 samples). This method is known to perform well with Spearman's correlations regardless of distribution type, even when the sample size is very small (Ruscio, 2008). All descriptive statistics and statistical tests were calculated in SPSS Version 29.0.

Results

Descriptive Findings

Table 1 displays all articles that met inclusion for the literature review and summarizes the descriptive findings. These include publication source, sample size, outcomes predicted, whether the model was applied or theoretical, the type of data source used, whether or not the team was cross-disciplinary or incorporated PD, the data source limitations acknowledged by the authors, the classification model and test-train approaches used, and the performance of the best performing models for each study. We elaborate on these findings in the subsections below.

Table 1 Sources Included in the Systematic Review of Sophisticated Predictive and Prescriptive Analytics in Child Welfare

Author(s) (date). Publication Source, Location (*N*)	Outcome Predicted (Purpose of Algorithm); Data Source	(a) Cross-Disciplinary Collaboration (b) Participatory Design	Acknowledgment of Limitations	Classification Model; Performance of Best Performing Model; Training/Testing Approach
Amrit et al. (2017). CS journal, Netherlands (*N* = 13,170)	Predict presumed maltreatment (applied); Medical data	(a) No - CS/ML only (b) Yes - PD	None	**TB***, NLC (Naive Bayes, **random forest [RF],*** Support vector machine [SVM]); AUC = 0.91 (excellent); k-fold cross-validation
Benesh (2017). Published dissertation, US (*N* = 727)	Predict type of placement setting and placement changes 18 months post-initial placement (theoretical); NSCAW I data	(a) No - HHS only (b) No - PD	Disparities	TB (RF); RQ1: κ = 0.35, MCC = 0.38 (fair); RQ2: κ = 0.16, MCC = 0.10 (poor); holdout method (60/40 split)
Camasso and Jagannathan (2019). SS journal, US (*N* = 52)	Predict state-level child maltreatment fatalities (theoretical); 1992–2013 NCANDS data	(a) No - SDA only (b) No - PD	None	Linear (parametric dynamic regression); Specificity and sensitivity = 60–75% (No rating); train/test approach unknown
Chouldechova et al. (2018). Published proceedings, US (*N* = 31,438)	Predict removal from home within two years of screened-in call (applied); Public child welfare agency dataset	(a) Yes - CS/ML, HHS, SDA (b) Yes - PD	Disparities, proxy variables, surveillance	**TB***, Linear, NLC (logistic regression, SVM, RF, **XGBoost***); AUC = 0.80 (good); holdout method (70/30 split)
Daley et al. (2016). SS journal, US (*N* = 64,126)	Predict maltreatment substantiation (theoretical); Public child welfare agency, CJS, and demographic datasets and private marketing research data	(a) Yes - HHS, CJS, SDA (b) Yes - PD	None	Linear (negative binomial type II regression model); Areas labeled as high-risk predicted 98% of observed maltreatment cases (No rating); validated by overlaying risk terrain map with map of observed child maltreatment cases
Elgin (2018). SS journal, US (*N* = 233,633)	Predict failure to attain legal permanency (theoretical); 2013 AFCARS data	(a) No - SDA (b) No - PD	None	**TB***, linear, **NLC*** (logistic regression, partial least squares discriminant analysis, Elastic Net/Lasso, **neural networks***, SVM, multivariate adaptive regression splines, classification trees, boosted trees, **RF***); AUC = 0.99, MCC = 0.87, κ = 0.87 (excellent); holdout method (75/25 split) with k-fold cross-validation during the training phase for hyperparameter optimization
Horikawa et al. (2016). SS journal, Japan (*N* = 716)	Predict maltreatment substantiation recurrence within one year of initial substantiation (theoretical); Public child welfare agency dataset	(a) No - SDA (b) Yes - PD	Subjectivity	Linear (stepwise multiple logistic regression); AUC = 0.69 (poor); Training data same as test data
Jolley (2012). Published dissertation, US (*N* = 6,747)	Predict maltreatment substantiation recurrence (theoretical); Public child welfare, SS, CJS, JJS, and demographic datasets	(a) No - HHS (b) No - PD	None	TB, Linear, **NLC*** (RF, **neural network***, logistic regression); AUC = 0.81, MCC = 0.51, κ = 0.51 (good); k-fold cross-validation
Rodriguez et al. (2019). CS journal, US; (*N* = 12,017)	Predict unsubstantiated maltreatment (theoretical); 2017 NCANDS data	(a) No - HHS (b) No - PD	Disparities	TB (RF); MCC = 0.46, κ = 0.43 (fair); holdout method (70/30 split per personal communication with author)
Schwartz et al. (2017). SS journal, US (*N* = 78,394)	Predict (1) maltreatment substantiation and (2) type and intensity of services delivered (theoretical); Public and private child welfare agency and CJS datasets	(a) Yes - CS/ML, HHS (b) Yes - PD	Disparities, proxy variables, subjectivity	TB (decision trees [C5 and CHAID] with ensemble learning and boosting); (1) AUC = 0.87 (good), (2) AUC = 0.81 (good); train/test approach unknown
Thurston and Miyamoto (2018). SS journal, US (*N* = 700)	Predict serious maltreatment substantiation (theoretical); Public child welfare agency, CJS, and SS datasets	(a) No - HHS (b) No - PD	Disparities, proxy variables, surveillance	TB (model-based recursive partitioning [decision trees]); MCC = 0.11 – 0.19, κ = 0.10–0.16 (poor); train/test unknown
Vaithianathan et al. (2013). Medical journal, NZ (*N* = 57,986)	Predict maltreatment substantiation by age 5 (theoretical); Public child welfare agency and SS datasets	(a) Yes - CS/ML, HHS, SDA (b) No - PD	Proxy variables, surveillance	Linear (stepwise probit model); AUC = 0.76 (fair); holdout method (70/30 split)

Table 1 (continued)

Author(s) (date). Publication Source, Location (*N*)	Outcome Predicted (Purpose of Algorithm); Data Source	(a) Cross-Disciplinary Collaboration (b) Participatory Design	Acknowledgment of Limitations	Classification Model; Performance of Best Performing Model; Training/Testing Approach
Vaithianathan et al. (2018). Medical journal, NZ (*N* = 121,482)	Predict maltreatment substantiation by age 2 and estimate prevalence of injury or mortality by 3 (theoretical); Public child welfare agency, CJS, and SS datasets	(a) Yes - HHS, SDA (b) No - PD	Disparities, proxy variables, surveillance	Linear (logistic regression); AUC = 0.88 (good); train/test approach unknown
Walsh et al. (2020). SS journal, NZ (*N* = 3,883)	Predict odds of 2 + adverse childhood experiences by age 54 months (theoretical); Primary data collection	(a) No - SDA (b) No - PD	None	Linear (logistic regression); AUC = 0.76 (fair); holdout method (80/20 split)
Wilson et al. (2015). Medical journal, NZ (*N* = 62,273)	Predict maltreatment substantiation by age 2 (theoretical); Public child welfare agency, SS, CJS, and demographic datasets	(a) No - SDA (b) Yes - PD	Disparities	TB, **Linear***, NLC (gradient boosting, DMINE regression, neural networks, partial least squares, full logistic regression, **stepwise logistic regression***, logistic regression with backward elimination, decision trees, multilevel model); AUC = 0.87, MCC = 0.33–0.65, κ = 0.27–0.63 (good); holdout method (70/30 split)

Note. AFCARS = Adoption and Foster Care Analysis and Reporting System; AUC = Area under the receiver operating curve; CJS = Criminal justice system; CS = Computer science; CS/ML = Computer science and machine learning; HHS = Health and human services; JJS = juvenile justice system; MCC = Matthew's correlation coefficient; NCANDS = National Child Abuse and Neglect Data System; NLC = Non-linear classification model; NSCAW = National Survey of Child and Adolescent Well-being; PD = Participatory design; SDA = Social data analytics; SS = Social sciences; TB = Tree-based model

***Best performing models**

Project and Publication Characteristics

Of the 15 sources included within our review, most were published between 2016 and 2020 (80%, *n* = 12) in medical, social sciences, and computer science journals (80%, *n* = 12). One conference paper and two dissertations met criteria for the review. Most projects were conducted within the United States (60%, *n* = 9) or New Zealand (26.67%, *n* = 4) and studied maltreatment presumption (Amrit et al., 2017) or substantiation (Daley et al., 2016; Horikawa et al., 2016; Jolley, 2012; Schwartz et al., 2017; Vaithianathan et al., 2013; Wilson et al., 2015) (66.67%, *n* = 10), including maltreatment resulting in serious injury or death (Camasso & Jagannathan, 2019; Thurston & Miyamoto, 2018; Vaithianathan et al., 2018), as the outcome of interest. No outcomes other than maltreatment substantiation were tested until 2017, during and after which studies focused on adverse childhood experiences (Walsh et al., 2020) and factors that protected against maltreatment substantiation (Rodriguez et al., 2019), as well as failure to achieve legal permanency (Elgin, 2018), placement setting type, types of placement changes (Benesh, 2017), and likelihood of removal from the home (Chouldechova et al., 2018). Only one study focused on more than one outcome type: Schwartz et al. (2017) used two separate models to examine maltreatment substantiation and actions taken by the child welfare system (i.e., services recommended based on participants' characteristics and substantiation determination) as outcomes.

Thirteen (86.67%) sources offered theoretical perspectives on the use of predictive and prescriptive analytics in child welfare but did not attempt to apply their models to child welfare settings. The two applied projects were designed to guide decision-making in child welfare hotline (Chouldechova et al., 2018) and pediatric care settings (Amrit et al., 2017) based on predicted risk of child maltreatment. Both applied projects included organizational staff as co-investigators.

Methods Used

All but one source (Walsh et al., 2020) utilized secondary data analysis, which is typical for machine learning. Just under half (46.67%, *n* = 7) relied on data linked between child welfare agency and criminal or juvenile justice, public social services, or public demographic datasets, one of which linked these data to other data collected by a marketing research firm. The remaining studies involved analyses of federally mandated national datasets (26.67%, *n* = 4) and public child welfare agency (13.33%, *n* = 2) or medical (6.67%, *n* = 1) datasets with no linkages to other data sources. Most researchers (60%, *n* = 9) trained their models on sample sizes of at least 10,000, with two studies relying

on a sample size of over 100,000. Two studies based their analyses on sample sizes between 1,001 and 9,999 participants and four on a sample of fewer than 1,000 participants. Two of the above studies used U.S. districts, including the District of Columbia and Puerto Rico (Camasso & Jagannathan, 2019), and half-street blocks in Fort Worth, TX (Daley et al., 2016) as units of analysis. Regarding analytics, most studies tested only one algorithm (60%, $n = 9$), with a range of one to 12 algorithms total ($M = 3.33$; $Mdn = 1$). Ten studies (66.67%) tested classification models in only one of the three categories. Of these, four and six employed TB and linear models, respectively. Amrit et al. (2017) tested both TB and NLC models. The remaining four studies tested at least one model across each of the three categories.

Cross-Disciplinary Collaboration and Participatory Design

Five of the 15 sources (33.33%) were written by cross-disciplinary teams. Of the 10 that were not cross-disciplinary, one was written by representatives of a government agency focused on social data analytics (Wilson et al., 2015); two were dissertations written by PhD candidates in family and child sciences (Benesh, 2017) and social work (Jolley, 2012); one was authored solely by social work researchers (Rodriguez et al., 2019); one by computer science researchers (Amrit et al., 2017); three by economics and public policy researchers (Camasso & Jagannathan, 2019; Elgin, 2018; Walsh et al., 2020); and two by medical, public health, and nursing researchers (Horikawa et al., 2016; Thurston & Miyamoto, 2018).

Six of the 15 sources (40%) were labeled as PD. One of the articles was written by medical doctors, a project coordinator, and a decision-support analyst from a local pediatric medical center (Daley et al., 2016). Three additional sources were authored by representatives of the local child welfare system (Horikawa et al., 2016; Wilson et al., 2015), including an IT administrator at a public child welfare agency (Schwartz et al., 2017). The final two sources explicitly discussed their incorporation of PD principles. Amrit and colleagues (2017) indicated that they explained how their model worked to the health providers who would use it. Chouldechova et al. (2018) developed their model in collaboration with the child welfare system, meeting with several stakeholders throughout the process, including community members with histories of child welfare involvement.

Model Performance

AUC values for the ten sources that reported this metric are presented in Table 1. Of the 13 studies that received performance ratings, a little over half were rated as good to excellent (53.85%, $n = 7$), and just under half were rated as poor to fair (46.15%, $n = 6$). As for researchers' approaches for testing and training their models, most either used the holdout method (40%, $n = 6$) or did not report any method at all (26.67%, $n = 4$). Two additional studies (13.33%) used k-fold cross-validation. Of the remaining studies, one used a combination of the holdout method and k-fold cross-validation, one used the same data for testing and training their model, and one overlaid a map of substantiated child maltreatment cases with the map they created with their risk terrain model to validate their predictions.

Ethics, Equity, and Bias

We used both quantitative counts and a qualitative scan of the articles to explore the degree to which researchers considered issues of ethics, equity, and bias, especially in reference to concerns about how algorithms may contribute to inequity in child welfare (Gillingham, 2019a, b; Keddell, 2015; Saxena et al., 2020). Fewer than half (40%, $n = 6$) of research teams reported integrating considerations for equity, ethics, or bias into model development or implementation. Only three studies (20%) reported evaluating model performance relative to ethics, equity, and bias. Below, we describe the design, implementation, or evaluation methods used in these studies.

First, in Chouldechova et al.'s (2018) study, the research team consulted with an independent evaluator who provided equity-related recommendations based on a formal review of the project. For example, the consultant approved the inclusion of race/ethnicity as a predictor in the model only if it enhanced model performance and advised the county not to disclose predicted risk scores to workers investigating screened-in calls to prevent bias from being introduced into investigations. They also evaluated their models for racial equity by comparing AUC across racial subgroups to determine whether models performed better for one race versus another, and inspected placement rates relative to predicted risk across races to evaluate the degree to which models over or underestimated risk based on race. Finally, they evaluated bias among child maltreatment hotline workers by comparing mandatory override rates across risk levels. Rates were similar regardless of risk, suggesting that decisions were influenced more by workers' subjective assessments than by the algorithm's recommendations.

Similarly, Wilson et al. (2015) evaluated equity by comparing the rates at which their model identified Māori children as at-risk for maltreatment to observed rates of substantiation and found that the algorithm's predictions were out of proportion to the percentage of Māori children within the maltreated population. The researchers attempted to correct this problem by building two separate algorithms:

one for Māori children and one for other racial/ethnic subgroups. This approach did not achieve the desired effect.

Schwartz et al. (2017) evaluated unethical decision-making practices in the child welfare system, defined as child welfare referrals that resulted in unfounded allegations. They found that 40% of referrals were unfounded based on data collected from hotline workers and investigators. Unfounded referrals increased the risk of repeat child welfare involvement to 175%. In comparison, their algorithm was able to identify cases that should be referred to court and to services requiring court involvement at accuracy rates of 90% and 93%, respectively – a substantial improvement over the 60% accuracy of the decision-making process in use at the time of the study. They concluded that adopting their model as a decision-making tool could substantially lower the risk of an unsubstantiated referral among families reported for potential maltreatment. These researchers did not appear to integrate considerations for ethics into their model's initial design; instead, they evaluated its potential for aiding ethical decision-making.

The remaining four research teams (26.67%) strove to minimize inequitable or unethical predictions without necessarily evaluating their efforts to do so. In particular, Rodriguez et al. (2019) identified surveillance and stigmatization of Black, Indigenous, and People of Color and low-income families within the child welfare system as their underlying premise for designing an algorithm that predicted unsubstantiated maltreatment based on protective rather than risk factors. Walsh et al. (2020) added protective factors to their model, but did not identify racial or socioeconomic equity as the motive. Jolley (2012) criticized actuarial risk assessments for relying on assumptions of linearity that obscure complex relationships and result in predictions that over or underestimate the risk of child maltreatment at unacceptably high rates. These ethical concerns guided her decision not to assume linear relationships between variables. Finally, Vaithianathan et al. (2013) excluded race/ethnicity from their model due to concerns about reinforcing racial stereotypes.

Synthesis of Results

The results of the Mann-Whitney U tests ($N = 13$) examining differences in performance ratings of best performing algorithms based on authors' dichotomous disciplinary affiliations (i.e., CS/ML, SDA, HHS), the cross-disciplinarity of the team, and their use of PD are presented in Table 2. None of these tests were statistically significant, indicating that there were no differences when comparing the distributions of performance ratings for any two groups. The Spearman's rank tests ($N = 13$) between performance rating and the number of model categories tested, $r(11) = 0.70$, $p = .008$, 95% CI [0.44, 0.85], and total number of algorithms tested, $r(11) = 0.81$, $p < .001$, 95% CI [0.62, 0.93], were positive, statistically significant, and exhibited large effect sizes. In other words, the larger the number of model categories or algorithms tested, the higher the performance rating of the best performing model. Performance rating and sample size were also positively correlated, $r(11) = 0.77$, $p = .002$, 95% CI [0.33, 0.96] and yielded a strong effect, suggesting that as sample size increased, so did the performance rating of the best performing algorithm.

The results of the Mann-Whitney U tests ($N = 15$) examining differences in the number of data source limitations acknowledged by the authors based on disciplinary affiliation, cross-disciplinarity of the team, and their use of PD are also presented in Table 2. None of these tests were statistically significant, indicating that there were no differences between the distributions of data source limitations for any two groups. Finally, Table 2 presents the findings of the Mann-Whitney U tests comparing group differences on the degree to which they integrated considerations for ethics, equity, and bias into the design/implementation or evaluation of their models. We used the same groupings as in the previous two sets of Mann-Whitney U tests ($N = 15$) and found no statistically significant differences between the distributions of any of the two groups. In sum, factors related to disciplinary affiliation, cross-disciplinarity, and PD did not seem to influence how teams addressed data source limitations or ethics, equity, and bias in the design, implementation, or evaluation of their models.

Discussion

This systematic literature review explored the state of the literature on machine learning as a utility for child welfare decision-making. We addressed four research questions: (1) How and where is algorithmic decision-making being used in child welfare? (2) To what extent are these projects cross-disciplinary or participatory? (3) To what extent do scholars address ethics, equity, and bias in their reporting of data source limitations, algorithmic design, and model implementation and performance? and (4) What factors contribute to the performance of models and the degree to which scholars address ethics, equity, and bias in their reporting of data source limitations, algorithmic design, and model implementation and performance?

Summary and Contextualization of Findings

We identified 15 articles that used algorithms for child welfare decision-making, all of which predicted an outcome based on independent variables (i.e., supervised machine learning). Studies primarily predicted risk of maltreatment

Table 2 Mann-Whitney U tests

Outcome Variable	Groups	*Mdn*	*M Rank*	*U*	*z*	*p*
1. Performance rating ($N = 13$)	CS/ML[a]					
	Yes ($n = 4$)	3.0	8.88	10.50	-1.206	0.269
	No ($n = 9$)	2.0	6.17			
	SDA					
	Yes ($n = 7$)	3.0	7.36	18.50	-0.372	0.788
	No ($n = 6$)	2.5	6.58			
	HHS					
	Yes ($n = 8$)	2.5	6.25	14.00	-0.916	0.382
	No ($n = 5$)	3.0	8.20			
	Cross-disciplinary[a]					
	Yes ($n = 4$)	3.0	8.00	14.00	-0.604	0.548
	No ($n = 9$)	2.0	6.56			
	PD					
	Yes ($n = 5$)	3.0	8.30	26.50	-0.992	0.354
	No ($n = 8$)	2.0	6.19			
2. Acknowledgment of data source limitations ($N = 15$)	CS/ML[a]					
	Yes ($n = 4$)	2.5	10.38	12.50	-1.306	0.205
	No ($n = 11$)	1.0	7.14			
	SDA[a]					
	Yes ($n = 9$)	1.0	7.67	24.00	-0.372	0.759
	No ($n = 6$)	1.0	8.50			
	HHS					
	Yes ($n = 9$)	2.0	9.89	10.00	-2.110	0.045
	No ($n = 6$)	0.0	5.17			
	Cross-disciplinary[a]					
	Yes ($n = 5$)	3.0	11.00	40.00	-1.935	0.064
	No ($n = 10$)	0.5	6.50			
	PD					
	Yes ($n = 6$)	1.0	8.50	24.00	-0.372	0.759
	No ($n = 9$)	1.0	7.67			
3. Ethics, equity, and bias ($N = 15$)	CS/ML[a]					
	Yes ($n = 4$)	1.0	10.13	13.50	-1.218	0.188
	No ($n = 11$)	0.0	7.23			
	SDA					
	Yes ($n = 9$)	0.0	8.00	27.00	0.000	1.00
	No ($n = 6$)	0.5	8.00			
	HHS					
	Yes ($n = 9$)	1.0	8.67	21.00	-0.776	0.516
	No ($n = 6$)	0.0	7.00			
	Cross-disciplinary[a]					
	Yes ($n = 5$)	1.0	9.00	30.00	-0.672	0.571
	No ($n = 10$)	0.0	7.50			
	PD					
	Yes ($n = 6$)	0.5	8.75	22.50	-0.582	0.662
	No ($n = 9$)	0.0	7.50			

Note. Exact two-tailed p-values were used for all tests given the small sample size and exploratory nature of the analysis. All p-values were corrected for ties. The Bonferroni adjusted p-value for $p < .05$, calculated for each of the three sets of tests separately, is $p < .01$. CS/ML = Computer science and machine learning; HHS = Health and human services; PD = Participatory design; SDA = Social data analytics

[a]This result should be interpreted with caution due to the small size of one sample relative to the other

substantiation, the likelihood of achieving permanency, and placement changes for youth in foster care. Most were conducted in the United States within the last five years using secondary data and employed algorithms that achieved good to excellent performance.

Notably, studies varied on the extent to which they used methods to avoid overfitting, which may have led to artificially inflated performance metrics. Further, three of the studies that achieved poor or fair performance (Benesh, 2017; Horikawa et al., 2016; Thurston & Miyamoto, 2018) had sample sizes of less than 1000. Data of these sizes are smaller than typical machine learning datasets, which can be trained on millions or billions of cases. Large datasets are most important for nonlinear algorithms (e.g. neural networks), which benefit from being more flexible with the downside of requiring more training data (Sen, 2021). Thus, studies that trained nonlinear algorithms on smaller datasets (e.g., Benesh 2017) could have lacked sufficient training data to accurately predict the outcome. Alternatively, models that performed poorly could have been underfit, which would suggest the need for more training time, independent variables, or specificity. The findings of our statistical tests align with the widely held conception that models trained on larger sample sizes are better able to optimize performance.

It is important not to confuse performance with ethical, equitable, or non-biased algorithms. Indeed, if data are patterned in a way that embeds existing biases, the algorithm may accurately predict these biases, ultimately reproducing institutionalized inequalities (Gillingham, 2019b). Researchers seeking to avoid bias may choose strategies that reduce the accuracy of their models but increase fairness (e.g., optimizing for racial equity). Only six sources reported integrating considerations for equity, ethics, or bias into model development or implementation, and only three evaluated performance relative to these indicators, among which Chouldechova et al.'s (2018) was the most comprehensive and Jolley's (2012) and Vaithianathan et al.'s (2013) the most limited.

Notably, the three sources that examined equity, ethics, and bias included authors from HHS disciplines, especially social workers. Social work was established in the United States in the late 19th century in response to major child welfare, public and mental health, housing, and labor movements that demanded ethical and competent methods for alleviating social problems like poverty and child maltreatment (Trattner, 1999). The International Federation of Social Workers (2014) defines social work as a "discipline that promotes social change and development, social cohesion, and the empowerment and liberation of people." This mission informs how social workers are educated. For instance, in the United States, the core values of the profession—social justice, service, dignity and worth of the person, integrity, and competence—are woven into the National Association of Social Workers' (2021) professional *Code of Ethics* and the Council on Social Work Education's (2022) *Educational Policy and Accreditation Standards for Baccalaureate and Master's Social Work Programs*.

This is not the case for most other professions, especially CS/ML, for which ethics courses are often not made mandatory in the curriculum (Fiesler et al., 2020). Only recently has there been a surge of CS/ML studies that consider bias, ethics, and equity (e.g., racism in technology) (Ogbonnaya-Ogburu et al., 2020). Although our findings did not point to any group differences between projects with and without HHS scholars or between cross-disciplinary and monodisciplinary teams, future studies should explore how differences in education and training between disciplines translate to differences in the handling of ethics, equity, and bias in the design and evaluation of machine learning algorithms.

Notwithstanding failures to evaluate their models for ethics, equity, or bias, four research teams did demonstrate commitments to minimizing inequitable or unethical predictions that are known to disproportionately harm historically oppressed groups. Aware of the racialized and class-based stigma associated with the tendency to over-focus on risk, Rodriguez et al. (2019) designed an algorithm that used protective factors to predict the positive outcome of *unsubstantiated* maltreatment. Walsh et al. (2020) utilized a more conservative approach to challenging the child welfare system's heavy reliance on risk detection to identify and prevent child maltreatment; instead of forgoing risk prediction altogether, they added protective factors to their model. However, this decision appeared to be driven more by a basic concern for ethics than an aspiration toward achieving racial and class equity. To enhance equity, scholars should consider applying asset-based approaches such as these. Similar to the strengths-based approach commonly used in social work practice, asset-based approaches focus on strengths and view diversity, culture, and sociodemographic characteristics as positive assets (Pattison et al., 2022). Everyone involved—designers, service users, and social workers—are valued for their capacities instead of characterized by what they lack or need to change (Pattison et al., 2022).

None of the remaining studies addressed issues of equity or bias. Even when these concerns were identified, most did not attempt to ameliorate them by adjusting the algorithmic formula. Although one Mann-Whitney U test shown in Table 2 initially pointed to a potential difference in the degree to which authors discussed data source limitations related to ethics, equity, and bias based on whether or not HHS scholars were on the team, this statistic was rendered non-significant after adjusting the p-value for multiple comparisons. Moreover, factors related to disciplinary affiliation, cross-disciplinarity, and PD did not appear to influence

whether methodological tools were used to address these concerns. This null finding could be related to the limitations of the systematic literature review described below (e.g., lack of statistical power), the multifaceted and complex nature of inequity and bias, or the novelty of this area of research. The CS/ML literature has only recently begun to examine bias, equity, and ethics in algorithms (Ogbonnaya-Ogburu et al., 2020). Perhaps we do not fully understand the ways bias and inequity can occur in algorithm design and thus, have limited ways to address it.

Limitations

This systematic literature review is subject to several limitations. First, we only reviewed scholarly sources. White papers (e.g., Blatt et al., 2016), news articles (e.g., Osher, 2018), and agency reports have also documented the increasing use of child welfare algorithms in recent years. These projects often do not appear in the academic literature and are not subject to the same external scrutiny like those included within this review. When child welfare agencies must rely on private contractors who have profit motives for technical expertise, the resulting algorithms may receive less internal scrutiny, as well. In some cases, the private contractor may have the right to limit access to information about how the algorithm works, resulting in a lack of transparency (Brauneis & Goodman, 2018). Granting private corporations ownership over tools used to administer public services reduces public accountability and gives corporations control over how policy is enforced on the ground, a phenomenon that Brauneis and Goodman (2018) have referred to as "policy outsourcing" (p. 111). Some privately developed algorithms launched in child welfare agencies were later removed due to performance issues or concerns with the private contractor. For example, Eckerd Kids' Rapid Safety Feedback system was abandoned in Illinois after the algorithm greatly overestimated risk (Gillingham, 2019a). Our review did not examine these particular cases.

Second, only one author was primarily responsible for extracting and coding the data items, albeit with consistent feedback and guidance from the second author, so we did not calculate interrater agreement between multiple independent coders as is customarily recommended. However, both of the first two authors reviewed all 15 sources several times and frequently expressed and resolved disagreements during meetings and via Google sheet comments. The coding and data extraction processes were iterative and involved several in-depth discussions. Third, although we chose methods that would theoretically maximize power while minimizing Type I error based on the properties of our data, it is doubtful our small sample size afforded us sufficient power to observe true effects. Further, the high level of heterogeneity between studies and possibility that research teams engaged in certain activities without reporting them (e.g., PD) limits the extent to which we can draw comparisons at all. The results of our statistical tests should therefore be interpreted with caution.

Fourth, because we relied on authorship lists to determine cross-disciplinarity and domain expertise and efforts to gather additional details from corresponding authors were unsuccessful, we may have overlooked the presence of non-author participants. The discipline of Human Computer Interaction (HCI) offers growing attention to the need for domain experts, especially in public-impact research (Stege & Breitner, 2020; Wang et al., 2020; Weerts et al., 2019). However, HCI often treats domain experts as consultants, rather than full partners who inform research questions, data cleaning, and implementation (e.g., Lupton, 2017; Milton et al., 2021; VanHeerwaarden et al., 2018). More research is needed regarding the impact of domain experts on the use of algorithms in child welfare and the strategies child welfare agencies can use to improve partnerships with computer scientists.

Finally, investigating how the performance of machine learning models may compare to those of other decision-making methods in child welfare was not feasible given the small number of sources available and heterogeneity across studies in outcomes, predictors, data sources, and analytic methods. As the body of literature in this area grows, future research might expand on other meta-analyses that have compared child welfare decision-making instruments on performance metrics. For instance, a review and meta-analysis by van der Put et al. (2017) identified 30 studies that assessed the validity of actuarial, consensus-based, and structured clinical judgment-based instruments to predict future child maltreatment. They reported an overall AUC based on 67 values derived from these studies (AUC = 0.70), as well as AUCs for each instrument type, finding that actuarial assessments outperformed both consensus-based and structured clinical judgment-based instruments. Future reviews might pool the effect sizes of groups of homogeneous studies and compare them to those of these other tools.

Future Directions

The child welfare system has long worked to address concerns about ethics and professionalism. Given the complex nature of child welfare and the potential for algorithms to greatly impact the lives of children, youth, and families, it is imperative that future efforts are led by domain experts and informed by service recipients. However, child welfare systems must build more internal knowledge of this type

of innovation. One possible solution to the lack of computer science expertise within child welfare is to extend Title IV-E workforce funds to grow partnerships between child welfare scholars, data scientists, and agency administrators (Griffiths et al., 2018; Zlotnik, 2003). Community-university partnerships between child welfare agencies and iSchools or CS/ML programs might also achieve this goal. These partnerships could offer internships to students in CS or health informatics to help design algorithms and train child welfare practitioners.

Often, computer science research is concerned with the ability to predict outcomes but not the implications of applied prediction. The partnerships we recommend may help inform the practical implications of model performance. For instance, accuracy that looks good in a research paper (e.g., 75% accuracy) does not look nearly as helpful when applied to real-world decision-making (Gillingham, 2019a). The many concerns raised about the use of algorithms in child welfare do not appear to have stopped vendors from attempting to bring algorithmic decision-making tools to the marketplace. Child welfare agencies are eager to find solutions that solve problems of high staff turnover, underfunding, workload, and fairness, and on the surface, data-driven solutions seem to answer that call. Algorithms may alleviate some of these issues, particularly if the data are collected for the purposes of modeling, domain experts are full partners on the project, and findings point to reforms that could improve the child welfare system. Overall, our current work demonstrates that these aspects of projects are not being adequately documented in the academic literature and that a uniform reporting protocol may be needed to guide this area of research as it continues to grow.

Acknowledgements This project was funded by the National Science Foundation (NSF Award #1939579).

Declarations

Conflicts of Interest The authors report that they have no conflicts of interest to disclose.

Ethics approval This research does not involve human participants and therefore, no other concerns for ethical standards are applicable.

References

Abramovitz, M., & Zelnick, J. (2015). Privatization in the human services: Implications for direct practice. *Clinical Social Work Journal*, *43*(3), 283–293. https://doi.org/10.1007/s10615-015-0546-1

Adusah, A. K., & Brooks, G. P. (2011). Type I error inflation of the separate-variances Welch t test with very small sample sizes when assumptions are met. *Journal of Modern Applied Statistical Methods*, *10*(1), 362–372. https://doi.org/10.22237/jmasm/1304224320

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, *88*, 402–418. https://doi.org/10.1016/j.eswa.2017.06.035

Baird, C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk assessment in child protective services: Consensus and actuarial model reliability. *Child Welfare*, *78*(6), 723–748.

Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential? *Vikalpa: The Journal for Decision Makers*, *38*(4), 1–12. https://doi.org/10.1177/0256090920130401

Benesh, A. S. (2017). *Predicting child welfare future placements for foster youth: An application of statistical learning to child welfare* (Publication No. 10258386) [Doctoral dissertation, Florida State University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/predicting-child-welfare-future-placements-foster/docview/1915941568/se-2

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Blatt, E., Clanton, S., Duggan, M., & Mann, J. (2016). *From automated to comprehensive: What child welfare organizations need to succeed* (White Paper No. ZZW03399-USEN-00). International Business Machines Watson Health for Social Programs. https://www.ibm.com/downloads/cas/N4QLPQEZ

Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *The Yale Journal of Law & Technology*, *20*(1–3), 103–176. https://digitalcommons.law.yale.edu/yjolt/vol20/iss1/3

Brindley, M., Heyes, J., & Booker, D. (2018). Can machine learning create an advocate for foster youth? *Journal of Technology in Human Services*, *36*(1), 31–36. https://doi.org/10.1080/15228835.2017.1416513

Camasso, M. J., & Jagannathan, R. (2019). Conceptualizing and testing the vicious cycle in child protective services: The critical role played by child maltreatment fatalities. *Children and Youth Services Review*, *103*, 178–189. https://doi.org/10.1016/j.childyouth.2019.05.024

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness Accountability and Transparency*, *81*, 134–148. http://proceedings.mlr.press/v81/chouldechova18a/chouldechova18a.pdf

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. https://doi.org/10.1111/jfcj.12086

Council on Social Work Education (2022). *Educational policy and accreditation standards for Baccalaureate and Master's social work programs*https://www.cswe.org/getmedia/94471c42-13b8-493b-9041-b30f48533d64/2022-EPAS.pdf

D'andrade, A., Austin, M. J., & Benton, A. (2008). Risk and safety assessment in child welfare: Instrument comparisons. *Journal of Evidence-Based Social Work*, *5*(1–2), 31–56. https://doi.org/10.1300/J394v05n01_03

Daley, D., Bachmann, M., Bachmann, B. A., Pedigo, C., Bui, M. T., & Coffman, J. (2016). Risk terrain modeling predicts child maltreatment. *Child Abuse & Neglect*, *62*, 29–38. https://doi.org/10.1016/j.chiabu.2016.09.014

Dare, T., & Gambrill, E. (2017). *Ethical analysis: Predictive risk models at call screening for Allegheny County* [Ethical Analysis]. Allegheny County Department of Human Services, Allegheny County Analytics. https://www.alleghenycountyanalytics.us/

wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf

Dataiku (2022). *Clustering (unsupervised ML)*. https://doc.dataiku.com/dss/latest/machine-learning/unsupervised/index.html

Elgin, D. J. (2018). Utilizing predictive modeling to enhance policy and practice through improved identification of at-risk clients: Predicting permanency for foster children. *Children and Youth Services Review*, *91*, 156–167. https://doi.org/10.1016/j.childyouth.2018.05.030

Elgin, D. J., & Carter, D. P. (2020). Higher performance with increased risk of undesirable outcomes: The dilemma of U.S. child welfare services privatization. *Public Management Review*, *22*(11), 1603–1623. https://doi.org/10.1080/14719037.2019.1637013

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fiesler, C., Garrett, N., & Beard, N. (2020). What do we teach when we teach tech ethics? A syllabi analysis. *SIGCSE '20: Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295. https://doi.org/10.1145/3328778.3366825

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, *33*(4), 111–117. https://doi.org/10.1215/07402775-3813015

Gillingham, P. (2019a). Can predictive algorithms assist decision-making in social work with children and families? *Child Abuse Review*, *28*(2), 114–126. https://doi.org/10.1002/car.2547

Gillingham, P. (2019b). Decision support systems, social justice and algorithmic accountability in social work: A new challenge. *Practice*, *31*(4), 277–290. https://doi.org/10.1080/09503153.2019.1575954

Gleeson, J. P. (1987). Implementing structured decision-making procedures at child welfare intake. *Child Welfare*, *66*(2), 101–112.

Griffiths, A., Royse, D., Piescher, K., & LaLiberte, T. (2018). Preparing child welfare practitioners: Implications for Title IV-E education and training partnerships. *Journal of Public Child Welfare*, *12*(3), 281–299. https://doi.org/10.1080/15548732.2017.1416325

Halimu, C., Kasem, A., & Newaz, S. H. S. (2019). Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing - ICMLSC 2019*, 1–6. https://doi.org/10.1145/3310986.3311023

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Horikawa, H., Suguimoto, S. P., Musumari, P. M., Techasrivichien, T., Ono-Kihara, M., & Kihara, M. (2016). Development of a prediction model for child maltreatment recurrence in Japan: A historical cohort study using data from a child guidance center. *Child Abuse & Neglect*, *59*, 55–65. https://doi.org/10.1016/j.chiabu.2016.07.008

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Huggins-Hoyt, K. Y., Briggs, H. E., Mowbray, O., & Allen, J. L. (2019). Privatization, racial disproportionality and disparity in child welfare: Outcomes for foster children of color. *Children and Youth Services Review*, *99*, 125–131. https://doi.org/10.1016/j.childyouth.2019.01.041

International Federation of Social Workers (2014, July). *Global definition of social work*. https://www.ifsw.org/what-is-social-work/global-definition-of-social-work/

Johnson, W., & L'Esperance, J. (1984). Predicting the recurrence of child abuse. *Social Work Research and Abstracts*, *20*(2), 21–26. https://doi.org/10.1093/swra/20.2.21

Jolley, J. M. (2012). *Applying neural network models to predict recurrent maltreatment in child welfare cases with static and dynamic risk factors* (UMI No. 3542505) [Doctoral dissertation, Washington University in St. Louis]. ProQuest Dissertations and Theses Global. https://www.proquest.com/dissertations-theses/applying-neural-network-models-predict-recurrent/docview/1152187850/se-2

Keddell, E. (2015). The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy*, *35*(1), 69–88. https://doi.org/10.1177/0261018314543224

Keddell, E. (2017). Comparing risk-averse and risk-friendly practitioners in child welfare decision-making: A mixed methods study. *Journal of Social Work Practice*, *31*(4), 411–429. https://doi.org/10.1080/02650533.2017.1394822

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lanier, P., Rodriguez, M., Verbiest, S., Bryant, K., Guan, T., & Zolotor, A. (2020). Preventing infant maltreatment with predictive analytics: Applying ethical principles to evidence-based child welfare policy. *Journal of Family Violence*, *35*(1), 1–13. https://doi.org/10.1007/s10896-019-00074-y

Lauritzen, C., Stein, A., & Fossum, S. (2018). Factors that determine decision making in child protection investigations: A review of the literature. *Child & Family Social Work*, *23*(4), 743–756. https://doi.org/10.1111/cfs.12446

Lupton, D. (2017). Digital health now and in the future: Findings from a participatory design stakeholder workshop. *Digital Health*, *3*, 1–17. https://doi.org/10.1177/2055207617740018

McElduff, F., Cortina-Borja, M., Chan, S. K., & Wade, A. (2010). When t-tests or Wilcoxon-Mann-Whitney tests won't do. *Advances in Physiology Education*, *34*(3), 128–133. https://doi.org/10.1152/advan.00017.2010

Milton, A. C., Hambleton, A., Dowling, M., Roberts, A. E., Davenport, T., & Hickie, I. (2021). Technology-enabled reform in a nontraditional mental health service for eating disorders: Participatory design study. *Journal of Medical Internet Research*, *23*(2), Article e19532. https://doi.org/10.2196/19532

Mitchell, T. M. (2006). *The discipline of machine learning* (CMU-ML-06-108). Carnegie Mellon University, School of Computer Science, Machine Learning Department. http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

National Association of Social Workers (2021). *Code of ethics of the National Association of Social Workers*https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English

Ogbonnaya-Ogburu, I. F., Smith, A. D., To, A., & Toyama, K. (2020). Critical race theory for HCI. *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376392

Osher, C. N. (2018, November 1). "This has to get fixed": Problems with $25.3 million upgrade to Colorado's child-protection computer system leave children at risk, officials say. *The Denver Post*. https://www.denverpost.com/2018/11/01/colorado-child-protection-computer-system-failing-children/

Pattison, S., Ramos Montañez, S., Svarovsky, G., & Tominey, S. (2022). *Engineering for equity: Exploring the intersection of*

engineering education, family learning, early childhood, and equity. TERC. https://blog.terc.edu/engineering-for-equity

Rodriguez, M. Y., DePanfilis, D., & Lanier, P. (2019). Bridging the gap: Social work insights for ethical algorithmic decision-making in human services. *IBM Journal of Research and Development*, *63*(4–5), 8:1–88. https://doi.org/10.1147/JRD.2019.2934047

Ruscio, J. (2008). Constructing confidence intervals for Spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods*, *7*(2), 416–434. https://doi.org/10.22237/jmasm/1225512360

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2020). A human-centered review of algorithms used within the U.S. child welfare system. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376229

Schwartz, I. M., York, P., Nowakowski-Sims, E., & Ramos-Hernandez, A. (2017). Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience. *Children and Youth Services Review*, *81*(C), 309–320. https://doi.org/10.1016/j.childyouth.2017.08.020

Sen, A. (2021, November 9). *Ensemble modeling for neural networks using large datasets – Simplified!* Analytics Vidhya Data Science Blogathon. https://www.analyticsvidhya.com/blog/2021/10/ensemble-modeling-for-neural-networks-using-large-datasets-simplified/

Shlonsky, A., & Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management. *Children and Youth Services Review*, *27*(4), 409–427. https://doi.org/10.1016/j.childyouth.2004.11.007

Simsek, A. S. (2023). The power and type I error of Wilcoxon-Mann-Whitney, Welch's t, and student's t tests for likert-type data. *International Journal of Assessment Tools in Education*, *10*(1), 114–128. https://doi.org/10.21449/ijate.1183622

Stege, N., & Breitner, M. H. (2020). *Hybrid intelligence with commonality plots: A first aid kit for domain experts and a translation device for data scientists* [Paper presentation]. 15th International Conference on Wirtschaftsinformatik, Potsdam, Germany. https://doi.org/10.30844/wi_2020_c7-stege

Thurston, H., & Miyamoto, S. (2018). The use of model based recursive partitioning as an analytic tool in child welfare. *Child Abuse & Neglect*, *79*, 293–301. https://doi.org/10.1016/j.chiabu.2018.02.012

Trattner, W. I. (1999). *From poor law to welfare state: A history of social welfare in America* (6th ed.). The Free Press.

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment. *American Journal of Preventive Medicine*, *45*(3), 354–359. https://doi.org/10.1016/j.amepre.2013.04.022

Vaithianathan, R., Rouland, B., & Putnam-Hornstein, E. (2018). Injury and mortality among children identified as at high risk of maltreatment. *Pediatrics*, *141*(2), https://doi.org/10.1542/peds.2017-2882. Article e20172882.

van der Put, C. E., Assink, M., & van Boekhout, N. F. (2017). Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments. *Child Abuse & Neglect*, *73*, 71–88. https://doi.org/10.1016/j.chiabu.2017.09.016

VanHeerwaarden, N., Ferguson, G., Abi-Jaoude, A., Johnson, A., Hollenberg, E., Chaim, G., Cleverley, K., Eysenbach, G., Henderson, J., & Levinson, A. (2018). The optimization of an eHealth solution (thought spot) with transition-aged youth in postsecondary settings: Participatory design research. *Journal of Medical Internet Research*, *20*(3), Article e8102. https://doi.org/10.2196/jmir.8102

Wald, M. S., & Woolverton, M. (1990). Risk assessment: The emperor's new clothes? *Child Welfare*, *69*(6), 483–511. https://www.jstor.org/stable/45394134

Walsh, M. C., Joyce, S., Maloney, T., & Vaithianathan, R. (2020). Exploring the protective factors of children and families identified at highest risk of adverse childhood experiences by a predictive risk model: An analysis of the growing up in New Zealand cohort. *Children and Youth Services Review*, *108*, 104556. https://doi.org/10.1016/j.childyouth.2019.104556

Wang, T., Gu, H., Wu, Z., & Gao, J. (2020). Multi-source knowledge integration based on machine learning algorithms for domain ontology. *Neural Computing and Applications*, *32*(1), 235–245. https://doi.org/10.1007/s00521-018-3806-5

Weber, M., & Sawilowsky, S. (2009). Comparative power of the independent t, permutation t, and wilcoxon tests. *Journal of Modern Applied Statistical Methods*, *8*(1), 10–15. https://doi.org/10.22237/jmasm/1241136120

Weerts, H. J. P., van Ipenburg, W., & Pechenizkiy, M. (2019). Case-based reasoning for assisting domain experts in processing fraud alerts of black-box machine learning models. *Proceedings of KDD Workshop on Anomaly Detection in Finance (KDD-ADF '19)*. https://doi.org/10.48550/arXiv.1907.03334

Wilson, M. L., Tumen, S., Ota, R., & Simmers, A. G. (2015). Predictive modeling: Potential application in prevention services. *American Journal of Preventive Medicine*, *48*(5), 509–519. https://doi.org/10.1016/j.amepre.2014.12.003

Zlotnik, J. L. (2003). The use of Title IV-E training funds for social work education. *Journal of Human Behavior in the Social Environment*, *7*(1–2), 5–20. https://doi.org/10.1300/J137v07n01_02